Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)

Gilles Adda, Maxime Amblard, Karën Fort

6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) (JEP-TALN-RÉCITAL) [1]

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)

Gilles Adda, Maxime Amblard, Karën Fort (Éds.)

Nancy, France, 08-19 juin 2020

1. https://jep-taln2020.loria.fr/

Avec le soutien de

Message des présidents de l'AFCP et de l'ATALA

En ce printemps 2020, et les circonstances exceptionnelles qui l'accompagnent, c'est avec une émotion toute particulière que nous vous convions à la 6e édition conjointe des Journées d'Études sur la Parole (JEP), de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) et des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Après une première édition commune en 2002 (à Nancy, déjà !), et une expérience renouvelée avec succès en 2004, c'est désormais tous les quatre ans (Avignon 2008, Grenoble 2012) que se répète cet événement commun, attendu de pied ferme par les membres des deux communautés scientifiques voisines.

Cette édition 2020 est exceptionnelle, puisque dans le cadre des mesures sanitaires liées à la pandémie mondiale de COVID-19 (confinement strict, puis déconfinement progressif), la conférence ne peut avoir lieu à Nancy comme initialement prévu, mais se déroule à distance, sous forme virtuelle, soutenue par les technologies de l'information et de la communication. Nous remercions ici chaleureusement les organisateurs, Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni et Sylvain Pogodalla, qui ont dû faire preuve de souplesse, d'inventivité, de détermination, de puissance de travail, et de tant d'autres qualités encore, afin de maintenir la conférence dans ces circonstances, en proposant un format inédit. Grâce aux différentes solutions mises en œuvre dans un délai court, la publication des communications scientifiques est assurée, structurée, et les échanges scientifiques sont favorisés, même à distance.

Bien entendu, nous regrettons tous que cette réunion JEP-TALN-RECITAL ne permette pas, comme ses prédécesseurs, de nouer ou renforcer les liens sociaux entre les différents membres de nos communautés respectives – chercheurs, jeunes et moins jeunes, académiques et industriels, professionnels et étudiants – autour d'une passionnante discussion scientifique ou d'un mémorable événement social... Notre conviction est qu'il est indispensable de maintenir à l'avenir de tels lieux d'échanges dans le domaine francophone, afin bien sûr de permettre aux jeunes diplômés de venir présenter leurs travaux et poser leurs questions sans la barrière de la langue, mais aussi de dynamiser nos communautés, de renforcer les échanges et les collaborations, et d'ouvrir la discussion autour des enjeux d'avenir, qui questionnent plus que jamais la place de la science et des scientifiques dans notre société.

Lors de la précédente édition, nous nous interrogions sur les phénomènes et tendances liés à l'apprentissage profond et sur leurs impacts sur les domaines de la Parole et du TAL. Force est de constater que l'engouement pour ces approches dans nos domaines a permis un retour sur le devant de la scène des domaines liés à l'Intelligence Artificielle, animant parfois un débat tant philosophique que technique sur la place de la machine dans la société, notamment à travers le questionnement sur la vie privée de l'utilisateur. Ces questionnements impactent tant la Parole que le TAL, d'une part sur la place de la gestion des données, d'autre part sur les modèles eux-mêmes. Malgré ces questionnements, nous constatons que les acquis et les expertises perdurent, et les nouvelles approches liées à l'apprentissage profond ont permis un rapprochement des domaines de la Parole et du TAL, sans les dénaturer, à la manière des conférences JEP-TALN-RECITAL qui créent un espace plus grand d'échange et d'enrichissement réciproques.

Nous terminons ces quelques mots d'ouverture en remerciant l'ensemble des personnes qui ont rendu possible cet événement qui restera, nous l'espérons, riche et passionnant, malgré les circonstances. L'ATALA et l'AFCP tiennent tout d'abord à réitérer leurs remerciements aux organisateurs des JEP, de TALN et de RECITAL, qui sont parvenus à maintenir le cap à travers vents et marées. Nos remerciements vont également à l'ensemble des membres des comités de programme, dont le travail et l'implication ont permis de garantir la qualité et la cohérence du programme finalement retenu. Un grand merci aux relecteurs pour le temps et le soin qu'ils ont dédiés à ce travail anonyme et indispensable. Ils se reflètent dans la qualité des soumissions que chacun pourra découvrir sur le site de la conférence.

En conclusion, cette 6e édition conjointe JEP-TALN-RECITAL est exceptionnelle parce qu'elle se tient dans un contexte de crise généralisée — crise sanitaire, économique, voire sociale et politique. Mais nous

formons le vœu qu'elle reste également dans les annales pour la qualité des échanges scientifiques qu'elle aura suscités, et pour le message envoyé à nos communautés scientifiques et à la société dans son ensemble, un message de détermination et de confiance en l'avenir, où la science et les nouvelles technologies restent au service de l'humain.

Véronique Delvaux, présidente de l'Association Francophone de la Communication Parlée
Christophe Servan, président de l'Association pour le Traitement Automatique des LAngues

Préface

L'atelier ETeRNAL 2 — Éthique et TRaitemeNt Automatique des Langues — fait suite à une première édition qui s'est tenue à TALN en 2015. Il est soutenu par le projet OLKi, projet de recherche académique, financé par l'I-SITE Lorraine Université d'Excellence (PIA3). Les questions que nous souhaitions voir aborder concernent aussi bien les apports du TAL à l'éthique que nos responsabilités en tant que producteurs d'outils. Nous ne pouvons en effet pas faire semblant de ne pas savoir que ceux-ci rendent possibles des abus, des actes criminels, des violations des droits individuels.

Le Traitement Automatique des Langues et de la parole (TALP) est au cœur des enjeux éthiques du XXI^{e} siècle : accès aux données personnelles et protection de la vie privée, traitement (et croisement) des masses de données, délocalisation et production participative sont autant de problématiques qui sont en lien direct avec les applications que nous développons.

L'atelier est confronté comme tout le monde à la crise du Covid-19. Nous avons décidé de le maintenir et de faire vivre la thématique en utilisant d'autres modalités, avec des temps majoritairement asynchrones (présentation invitée, commentaires et réponses sur les articles), et des moments synchrones (réponses à des questions et expérience collaborative).

L'atelier souhaite fournir un espace de réflexion ouvert et interactif. Il rassemble cinq articles. Deux d'entre eux s'intéressent à la notion de biais et comment les identifier. Les trois autres ouvrent sur des problématiques moins directes pour notre champ disciplinaire soit parce que nos pratiques sont peu développées dans ce sens, par exemple avec la réplication d'expériences, soit parce que les questions sont émergentes, comme l'article sur les contenus que l'on retrouve sur le web. La démultiplication des outils et ressources mis à disposition nous interroge tant sur la possibilité de disposer des résultats que sur les contenus qu'on y trouve. Est-ce que les résultats de la science sont accessibles ? Est-ce que tout le web peut être considéré comme à disposition ? Le dernier article donne une perspective historique à une question primordiale, l'utilisation des résultats de la reconnaissance de la parole dans le monde judiciaire. Avons-nous réussi à faire entendre les limites de notre production scientifique ?

L'atelier est également une opportunité pour l'organisation d'une expérience collective sur l'évaluation des biais de perception. À partir d'une expérience accessible en ligne, Aurélie Névéol, CR CNRS au Limsi, a accepté d'organiser la passation collective des tests et de rendre accessible les résultats. Il ne s'agit pas tant de vérifier que nous sommes biaisés, nous le savons tous, mais de mesurer à quel point nous le sommes. L'expérience porte sur la mise en avant de processus inconscients généralement nourris par les clichés véhiculés par les stéréotypes. L'objectif est de faire prendre conscience de l'existence de ces biais au sein de notre communauté. L'expérience portera plus particulièrement sur la prise en compte du genre dans le contexte scientifique.

Par ailleurs, l'atelier a souhaité inviter Dirk Hovy, associate professor d'informatique à l'Université Bocconi à Milan. Dirk Hovy travaille sur les liens entre langue, société et apprentissage automatique. C'est en particulier un spécialiste des questions d'éthique pour le TALP. Dans son exposé, il reviendra sur l'influence prise par les systèmes automatiques depuis l'utilisation massive des réseaux de neurones. Ces outils restent massivement considérés comme des boîtes noires, sans que les biais qu'ils créent ou amplifient soient réellement pris en compte. Son exposé détaille le nouveau rôle dont les scientifiques du TALP ont hérité et qu'ils doivent maintenant tenir dans toute sa complexité.

Au travers de ces différentes modalités, articles, exposé et expérience collective, l'atelier ETeRNAL cherche à aborder la diversité de ces questions complexes, trop souvent mise de côté par la communauté scientifique pour des raisons compréhensibles.

Gilles Adda, Maxime Amblard et Karën Fort

Table des matières

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Pratiques d'évaluation en ASR et biais de performance

Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Nous proposons une réflexion sur les pratiques d'évaluation des systèmes de reconnaissance automatique de la parole (ASR). Après avoir défini la notion de discrimination d'un point de vue légal et la notion d'équité dans les systèmes d'intelligence artificielle, nous nous intéressons aux pratiques actuelles lors des grandes campagnes d'évaluation. Nous observons que la variabilité de la parole et plus particulièrement celle de l'individu n'est pas prise en compte dans les protocoles d'évaluation actuels rendant impossible l'étude de biais potentiels dans les systèmes.

ABSTRACT

Evaluation methodology in ASR and performance bias.

We propose a reflection on the evaluation practices of automatic speech recognition (ASR) systems. After defining the notion of discrimination from a legal point of view and the notion of equity in artificial intelligence systems, we look at the practices in large evaluation campaigns. Current protocols do not yet take into account the variability of speech, especially speaker variability, rendering the study of potential bias in systems impossible.

MOTS-CLÉS : reconnaissance automatique de la parole, évaluation, éthique.

KEYWORDS: automatic speech recognition, evaluation, ethics.

1 Introduction

Suite aux progrès amenés par l'essor combiné du big data et de l'apprentissage machine, les systèmes de TAL sont capables aujourd'hui d'atteindre des performances impressionnantes. Mais passé l'effervescence des premières réussites, un discours parallèle s'est construit sur l'impact de ces technologies sur nos sociétés (Boyd & Crawford, 2012; Barocas & Selbst, 2016; Hovy & Spruit, 2016). Une des études les plus médiatisées s'intéressant aux biais présents dans les systèmes issus d'apprentissages supervisés est celle rendue publique par Pro Publica dénonçant le système COMPAS (Angwin *et al.*, 2016). Ce système était utilisé par les cours de justice pour évaluer le taux de récidive d'une personne inculpée et présentait des résultats biaisés selon l'origine des individus. Par la suite, des biais ont également été découverts dans des systèmes de reconnaissance faciale (Buolamwini & Gebru, 2018) et de génération automatique de légendes d'images [1] ou encore de tri de CV. [2] Dans le domaine du

1. https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-p
Dernière consultation le 03/03/2020.
2. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
Dernière consultation le 13/03/2020

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 1–9.
HAL : hal-02750220.

TAL, les articles sur les biais de genre, notamment concernant les représentations vectorielles de mots (plongement de mots ou *word-embeddings*) et les systèmes de traduction automatique rappellent à la communauté le caractère hautement social et situé des données langagières (Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2017; Garg *et al.*, 2018; Vanmassenhove *et al.*, 2018). De manière assez surprenante en revanche, la littérature concernant l'existence possible de tels biais dans les systèmes de traitement automatique de parole reste pauvre. Cet article est une réflexion générale sur la notion d'équité dans la reconnaissance automatique de la parole [3] et sur l'utilisation du WER comme métrique d'évaluation. Il est organisé comme suit : une première partie présente la notion d'équité dans les performances de systèmes d'apprentissage automatique. Dans un second temps, nous présentons les pratiques actuelles d'évaluation des systèmes de reconnaissance automatique de la parole, en nous appuyant sur des grandes campagnes d'évaluation. Dans une troisième partie, nous questionnons les pratiques d'évaluation face à la variabilité des résultats.

2 Equité et systèmes d'apprentissage automatique

Les systèmes d'apprentissage automatique peuvent être résumés comme étant une modélisation algorithmique d'un processus décisionnel. D'un point de vue légal, Berendt & Preibusch (2014) distinguent la notion de différenciation de la notion de discrimination. Là où la différenciation se définit comme une distinction de traitement, et donc une prise de décision différente, selon un ensemble de caractéristiques ou de paramètres, la discrimination est une différenciation faite sur des caractéristiques considérées comme non-acceptables par le contrat social. En France, il existe 25 critères non-acceptés dont le sexe, l'identité de genre, les origines, la religion, la situation économique ou encore la situation familiale (LOI n° 2008-496 du 27 mai 2008 portant sur diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations [4]). Il est donc important de rappeler que les notions de discrimination et de biais restent fortement culturelles, la législation variant selon les pays. Un article de Sánchez-Monedero *et al.* (2020) s'intéressant aux systèmes automatiques pour l'embauche soulignait d'ailleurs que les travaux sur les biais de ces systèmes sont majoritairement faits en considérant le cadre socio-légal des Etats-Unis.

Comme souligné par Kate Crawford (2017) dans son intervention à NeurIPS, le terme biais, largement utilisé pour parler de systèmes discriminatoires, est polysémique et complique donc parfois les échanges entre les communautés de l'apprentissage machine et d'autres domaines comme le droit ou la linguistique. Si historiquement la notion de biais a un sens technique en statistiques, où il décrit les différences systématiques entre un échantillon et une population, il est aujourd'hui largement utilisé pour parler de discriminations, dont il est pratiquement devenu synonyme.

Face aux définitions légales, a émergé le concept d'équité (*fairness*) dans les systèmes automatiques. Ntoutsi *et al.* (2020) dénombrent plus de 20 définitions mathématiques différentes. Chen *et al.* (2018) distinguent deux types d'équité à savoir l'équité de groupe, et l'équité individuelle. L'équité individuelle pose l'hypothèse que pour des individus équivalents ne différant que par la valeur de la variable protégée, le résultat sera équivalent et se testera donc avec des modèles linéaires mixtes. L'équité de groupe suppose que les performances suivent des distributions similaires dans les sous-groupes créés par les différentes valeurs de la variable protégée et se mesure avec des tests statistiques comme le test U de Mann-Whitney.

3. Par la suite nous utiliserons l'acronyme anglais ASR qui signifie *Automatic Speech Recognition*
4. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000018877783

Afin de déterminer si le système ne donne pas les mêmes opportunités aux individus (*opportunity-based bias*) ou présente des différences significatives dans les résultats entre groupes (*outcome-based bias*), il est nécessaire d'avoir accès aux informations concernant la variable protégée pour quantifier ces biais. Ces informations sont cependant rarement prises en compte dans les pratiques d'évaluation de systèmes d'ASR. Si la récolte ou l'accès aux méta-données est un premier obstacle, d'une manière générale, la notion de variabilité intrinsèque à la parole semble disparaître dans les procédures d'évaluation.

3 Évaluation en ASR

3.1 Bref rappel de la genèse de la tâche

Historiquement, la première tâche de reconnaissance de la parole consistait à reconnaître les dix chiffres isolément pour un locuteur donné à l'aide d'un dispositif cablé (Davis *et al.*, 1952). À partir des années 1960, les méthodes numériques ont été introduites, améliorant les performances sur les mots isolés, la parole continue restant une tâche particulièrement complexe (Haton *et al.*, 2006). Les tâches ont donc d'abord été simplifiées en supprimant des facteurs de difficulté : reconnaissance de mots isolés, puis de mots enchaînés, souvent sur des configurations mono-locuteur. Les phénomènes de coarticulation présents sur de la parole continue et la variabilité inter-locuteurs ont été traités par la suite, grâce aux progrès en informatique et en électronique, pour permettre maintenant de traiter des situations de communication écologiques.

La reconnaissance de la parole ayant pour but principal d'accéder au message et donc au contenu lexical, tout ce qui relevait de la variation phonostylistique a été considéré comme du bruit. L'objectif était d'augmenter "la robustesse des systèmes à l'environnement (bruit, locuteurs...)" (Calliope, 1989). Ont donc été proposées des techniques comme la normalisation des paramètres acoustiques, permettant de gérer différents environnements (téléphone, radio, variabilité des microphones, etc.). Dans ce contexte, l'individu est considéré comme une source de variabilité de même que son sexe, son âge, son accent, sa catégorie sociale ou encore son état physique et émotionnel, chaque critère impactant la production de la parole (Kreiman & Sidtis, 2011). En évoluant vers des systèmes indépendants du locuteur, la variabilité due à l'individu a été prise en compte, d'abord à l'aide de modèles en fonction du genre, puis ensuite par les différentes techniques d'adaptation. Mais cette prise en compte de la variabilité des locuteurs et locutrices dans le développement des systèmes ne se retrouve pas dans les pratiques d'évaluation.

3.2 Évaluation

Lorsque sont reportés des résultats de systèmes de reconnaissance automatique de la parole, la métrique utilisée est le taux d'erreur-mots ou WER (*word-error rate*), basé sur la distance de Levenshtein et se calculant comme la somme des erreurs (insertion, délétion et substitution) de l'hypothèse divisée par le nombre de mots total dans la référence. En pratique, le WER est calculé à l'échelle du corpus de test, lissant ainsi les variations dues à la longueur des énoncés. Le développement des systèmes d'ASR ayant souvent donné lieu à des campagnes d'évaluation (campagnes NIST de 2002 à 2009 [5]

5. https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

pour l'anglais, campagnes ESTER [6] et ETAPE [7] pour le français), le report d'une mesure unique permettait la comparaison directe des systèmes entre eux, les données de test étant communes. En 2017, IBM reportait un WER de 5,5% sur SwitchBoard et 10,3% sur CallHome (Saon *et al.*, 2017), et Microsoft des WER de 5,1% et 9,8% sur ces mêmes corpus, considérant avoir atteint un niveau de performance similaire voire supérieur à l'humain (Xiong *et al.*, 2018). Mais ce mode d'évaluation de la parole, à l'aide d'une mesure unique, va de pair avec une conception complètement désincarnée du langage. On évalue le système en décorrélant complètement le fait que cette parole est produite de manière située, par un individu en contexte.

Comme expliqué dans la Section 3.1, la variabilité de la parole, venant des individus ou des environnements a été vue comme un bruit à gommer dans la conception des systèmes. Des campagnes ont donc vu le jour pour prendre en compte ces défis. Les différentes campagnes de NIST 2002 à 2009 ont notamment largement travaillé sur différents enjeux techniques dûs aux environnements sonores : bande-passante du téléphone, enregistrements bruités, séparation parole/musique, etc., et ces problématiques sont d'ailleurs toujours d'actualité avec des campagnes telles que CHiME. [8] La notion de variation stylistique est quant à elle prise en compte dans les campagnes MGB Challenge [9] où les performances sont reportées en fonction des différentes émissions avec des variations de WER pouvant aller jusqu'à plus de 30 points (Bell *et al.*, 2015). L'âge n'a pas encore été pris en compte dans des campagnes d'évaluation de grande envergure mais il est abordé dans certains travaux portant sur la reconnaissance de la parole des enfants (Kennedy *et al.*, 2017) ou la parole des personnes agées (Aman *et al.*, 2013) qui montrent des différences de performances liées à l'âge.

Notre travail de recherche s'intéresse plus spécifiquement à la variation de genre. Nous parlons de genre, car nous nous intéressons aux caractéristiques présentes dans la parole des individus et relevant d'une instanciation individuelle de la représentation sociale sexuée de l'individu (Ochs, 1992). En ASR, la plupart des études et des données langagières ne font référence qu'à deux catégories femme/homme. À notre connaissance, seules trois études se sont réellement intéressées aux différences de performances entre hommes et femmes dans des tâches d'ASR durant les deux dernières décennies. Adda-Decker & Lamel (2005) observaient de meilleures performances sur les voix de femmes. Les travaux plus récents de (Tatman, 2017), au contraire, mettaient en avant une performance moins bonne dans le sous-titrage automatique de YouTube sur les voix de femmes. Cependant, cette tendance n'était plus significative dans son étude de la même année avec Kasten, contrairement aux variations de performances en fonction de l'accent ou de l'origine éthnique (Tatman, 2017; Tatman & Kasten, 2017).

On observe donc que s'il est de coutume de reporter des résultats en fonction de certains types de variations considérées comme posant des problèmes techniques (environnement sonore, phonostyle), dans le cadre des campagnes, peu d'études se consacrent à la variation des performances en fonction des caractéristiques des individus. Le sexe, l'âge, l'appartenance ethnique, qui sont pourtant des variables protégées aux yeux de la loi française, ne sont pas prises en compte explicitement comme facteurs de variation des performances des systèmes. Or, s'assurer de l'équité de groupe pour ces facteurs constitue une étape nécessaire pour une diffusion éthique de ces technologies dans la société.

6. http://www.afcp-parole.org/camp_eval_systemes_transcription/present.html
7. http://www.afcp-parole.org/etape.html
8. https://chimechallenge.github.io/chime6/
9. http://www.mgb-challenge.org/

4 Une évaluation différenciée : première analyse sur les données de Librispeech

Dès les premiers systèmes d'ASR, les enjeux différenciés des évaluations en fonction des objectifs de chaques acteurs : chercheurs, industriels, etc., ont été problématisés (Pallett, 1985). En effet, évaluer les performances d'un système n'a de sens qu'au regard de l'usage qu'il en est fait. Cependant, les systèmes étant de plus en plus utilisés dans nos sociétés, il est nécessaire de penser l'impact qu'ils peuvent avoir sur la population. Dans le cadre de la reconnaissance automatique de la parole, l'évaluation en fonction des caractéristiques individuelles des locuteurs et locutrices n'est jamais abordée, alors que des outils d'évaluation permettent de prendre en compte ces préoccupations. Le National Institute of Standards and Technology [10] (NIST) a développé un outil d'évaluation des systèmes de reconnaissance automatique de la parole : le Speech Recognition Scoring Toolkit [11] (SCTK), largement utilisé et intégré dans la boîte à outils KALDI (Povey *et al.*, 2011). Cet outil permet des évaluations par locuteur ou locutrice, mais ces options sont rarement utilisées dans les reports de résultats.

Nous présentons ici un rapport d'évaluation d'un système d'ASR prenant en compte la variabilité propre à l'individu sur les données de Librispeech (Panayotov *et al.*, 2015). Le système utilisé a été développé en utilisant ESPnet (Watanabe *et al.*, 2018) et la recette fournie pour le corpus. Nous atteignons un WER de 4.2% sur le jeu de données nommé *test-clean*, validant ainsi notre système comme état de l'art.

La Figure 1 présente la répartition des WER par individu en fonction du genre, obtenue par notre système. Nous avons regroupé ici les deux partitions de test (clean et other) en raison du faible nombre de locuteurs et locutrices distinct·es (l'échantillon total contient 37 locutrices et 36 locuteurs). On semble observer une répartition différente en fonction du genre, mais un nombre plus conséquent d'observations nous permettrait d'avoir une distribution plus fine des WER pour attester ou non de l'équité de notre système. En l'état actuel des mesures, le test U de Mann-Whitney n'est pas significatif (U=715, p-valeur=0.59). Bien qu'indicatifs, ces résultats semblent montrer que les performances varient selon le genre des individus, ce que nous avions aussi observé dans une précédente étude sur le français (Garnerin *et al.*, 2019) et il serait intéressant d'étudier ces variations en fonction des différentes caractéristiques des individus, notamment dans les corpus largemement utilisés par la communauté.

5 Discussion

Dans une précédente étude (Garnerin *et al.*, 2019) nous avons montré que la faible présence des femmes dans les données médiatiques conduisait à un bias de performance genré. Face à ce constat, une solution serait de rééquilibrer les données, sacrifiant les performances du système pour assurer une équité de groupe. Nous ne cherchons pas à niveller par le bas, mais à alerter sur les pratiques actuelles d'évaluation des systèmes de reconnaissance de parole, qui ne permettent pas de connaître la variabilité de performance des systèmes sur différents groupes de locuteurs et locutrices. Cette évaluation lacunaire peut conduire à des problèmes éthiques dans la mesure où cela ne permet pas de définir

10. https://www.nist.gov/itl
11. https://github.com/usnistgov/SCTK

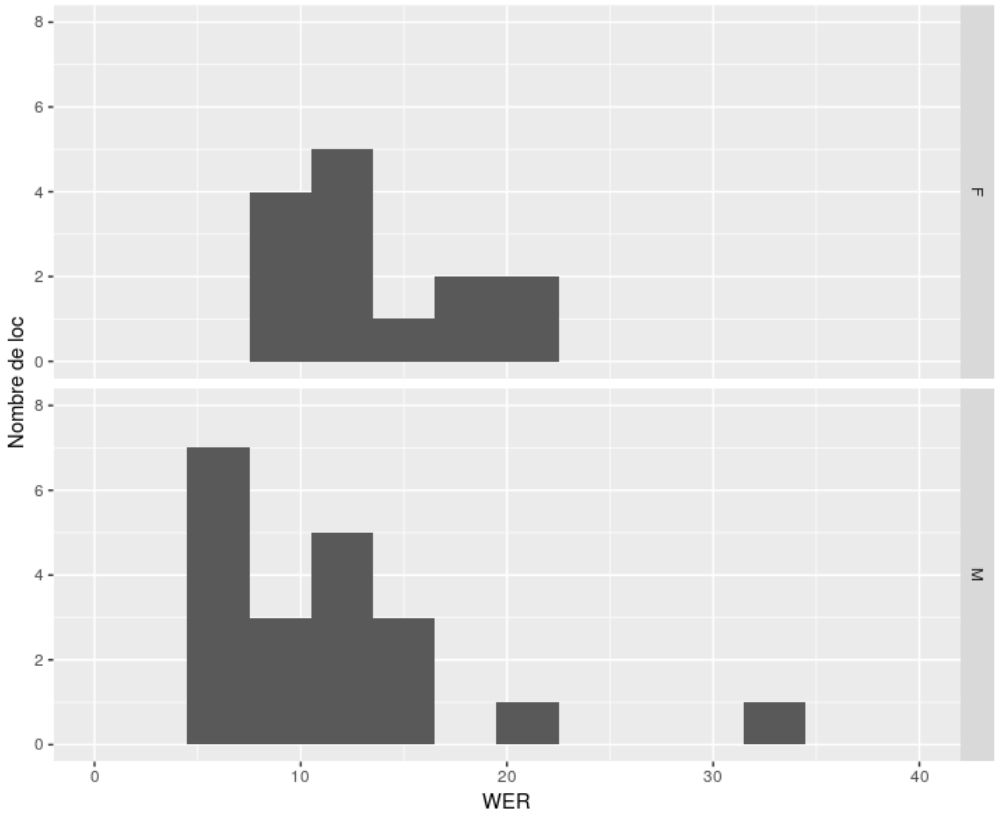


FIGURE 1 – Distribution des WER (en %) en fonction du genre (en haut : femmes ; en bas : hommes) sur le regroupement des jeux de données *test-clean* et *test-other* du corpus Librispeech

clairement les usages qu'il peut être fait des systèmes. En effet, la question éthique ne peut pas être envisagée indépendamment des usages. Actuellement, la reconnaissance de parole est principalement utilisée dans l'industrie pour un ensemble de tâches telles que le compte rendu de réunion ou la saisie de documents, mais également pour le sous-titrage et l'indexation de contenus audio-visuels ainsi que pour le maintien des personnes agées à domicile à l'aide de systèmes de domotique. La dictée vocale ou le compte-rendu ne posent pas en eux-même des problèmes particuliers, étant donné que ces systèmes fonctionnent en local, avec une étape d'adaptation à l'utilisateur ou à l'utilisatrice. En revanche, en ce qui concerne l'indexation et le sous-titrage, on peut se demander si cela ne va pas contribuer à l'invisibilisation de certaines catégories de personnes dans les médias : on peut penser aux femmes (Doukhan & Carrive, 2018), mais également à l'impact sur la parole accentuée. En ce qui concerne le service à la personne, la prise en compte du genre est incontournable étant donné que les hommes ne représentent que 38,9% de la population des 75 ans et plus, comme rapporté par l'édition 2019 du portrait social de l'INSEE. [12] Le faible nombre d'études sur les différences de performances entre hommes et femmes dans les systèmes de reconnaissance automatique de la parole s'explique donc peut-être par le faible impact sociétal des écarts de performances lors de l'utilisation actuelle de ces systèmes. En revanche, avec l'émergence des assistants vocaux, qui sont des services fonctionnant sur serveurs, sans adaptation directe au locuteur ou à la locutrice, on observe une volonté de faire de la voix la nouvelle interface de nombreux produits de service. On peut également se questionner sur l'effet que pourraient avoir des performances différenciées dans le cas de systèmes de traductions speech to text ou speech to speech.

12. https://www.insee.fr/fr/statistiques/4238375?sommaire=4238781#consulter

6 Conclusion

La notion d'éthique en traitement automatique des langues est un enjeu majeur dont la communauté de parole doit s'emparer. Dans cet article, nous proposons de repenser l'évaluation des systèmes de reconnaissance automatique de la parole en termes d'équité. Il est clair que la parole est un domaine dans lequel le sexe, l'identité de genre, l'âge, l'appartenance ethnique et la classe sociale sont des sources importantes de variabilité. Les possibilités d'analyse des performances des systèmes en fonction de ces variables protégées par la loi française existent. Nous ne pouvons que regretter qu'elles ne fassent pas partie des pratiques d'évaluation de la communauté. Les grandes campagnes d'évaluation ont pour objectif de comparer les systèmes des différentes équipes afin d'améliorer les architectures et ne comparent donc que les systèmes entre eux. Mais avec la diffusion de ces systèmes dans la société, il est nécessaire de penser une évaluation différente, dans laquelle il ne s'agit non pas que de trouver le meilleur système mais aussi de s'assurer de la conformité des systèmes au cadre légal, pour en limiter d'éventuels impacts sociétaux négatifs.

Références

ADDA-DECKER M. & LAMEL L. (2005). Do speech recognizers prefer female speakers ? In *Actes de INTERSPEECH 2005 (International Speech Communication Association)*, p. 2205–2208, Lisbon, Portugal : ISCA.

AMAN F., VACHER M., ROSSATO S. & PORTET F. (2013). Speech recognition of aged voice in the AAL context : Detection of distress sentences. In *Actes de SPED13 (Conference on Speech Technology and Human-Computer Dialogue)*, p. 1–8 : IEEE.

ANGWIN J., LARSON J., MATTU S. & KIRCHNER L. (2016). Machine bias. *ProPublica*, **23**.

BAROCAS S. & SELBST A. D. (2016). Big data's disparate impact. *California Law Review*, **104**, 671.

BELL P., GALES M. J., HAIN T., KILGOUR J., LANCHANTIN P., LIU X., MCPARLAND A., RENALS S., SAZ O., WESTER M. *et al.* (2015). The MGB challenge : Evaluating multi-genre broadcast media recognition. In *Actes de ASRU 2015 (Workshop on Automatic Speech Recognition and Understanding)*, p. 687–693 : IEEE.

BERENDT B. & PREIBUSCH S. (2014). Better decision support through exploratory discrimination-aware data mining : foundations and empirical evidence. *Artificial Intelligence and Law*, **22**(2), 175–209.

BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? Debiasing word embeddings. In *Actes de NeurIPS 2016 (Neural Information Processing Systems)*, p. 4349–4357.

BOYD D. & CRAWFORD K. (2012). Critical questions for big data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, **15**(5), 662–679.

BUOLAMWINI J. & GEBRU T. (2018). Gender shades : Intersectional accuracy disparities in commercial gender classification. In *Actes de FAT 2018 (Fairness, Accountability and Transparency)*, p. 77–91, New-York City, USA : ACM.

CALISKAN A., BRYSON J. J. & NARAYANAN A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

CALLIOPE (1989). *Ergonomie et évaluation du traitement de la parole par ordinateur*, In J. TUBACH, Éd., *La parole et son traitement automatique*, chapitre 26, p. 689–705. Paris : Masson.

CHEN L., MA R., HANNÁK A. & WILSON C. (2018). Investigating the impact of gender on rank in resume search engines. In *Actes de CHI 2018 (Conference on Human Factors in Computing Systems)*, p. 1–14, Montréal, QC, Canada.

CRAWFORD K. (2017). The trouble with bias. NIPS 2017 Keynote.

DAVIS K. H., BIDDULPH R. & BALASHEK S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, **24**(6), 637–642.

DOUKHAN D. & CARRIVE J. (2018). Description automatique du taux d'expression des femmes dans les flux télévisuels français. In *Actes de JEP 2018 (Journées d'Études sur la Parole)*, p. 496–504, Aix-en-Provence, France.

GARG N., SCHIEBINGER L., JURAFSKY D. & ZOU J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, **115**(16), E3635–E3644.

GARNERIN M., ROSSATO S. & BESACIER L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. In *Actes de AI4TV 2019 (Workshop on AI for Smart TV Content Production, Access and Delivery)*, p. 3–9, Nice, France : ACM. DOI : 10.1145/3347449.3357480.

HATON J.-P., CERISARA C., FOHR D., LAPRIE Y. & SMAÏLI K. (2006). *Introduction à la reconnaissance automatique de la parole*, In *Reconnaissance automatique de la parole : Du Signal à son Interprétation*, chapitre 1, p. 1–15. Paris : Dunod.

HOVY D. & SPRUIT S. L. (2016). The social impact of Natural Language Processing. In *Actes de ACL 2016 (Volume 2 : Short Papers)*, p. 591–598, Berlin, Allemagne : Association for Computational Linguistics. DOI : 10.18653/v1/P16-2096.

KENNEDY J., LEMAIGNAN S., MONTASSIER C., LAVALADE P., IRFAN B., PAPADOPOULOS F., SENFT E. & BELPAEME T. (2017). Child speech recognition in human-robot interaction : evaluations and recommendations. In *Actes de HRI 2017 (International Conference on Human-Robot Interaction)*, p. 82–90 : ACM/IEEE.

KREIMAN J. & SIDTIS D. (2011). *Physical Characteristics and the Voice : Can We Hear What a Speaker Looks Like ?*, In *Foundations of Voice Studies*, chapitre 4. Wiley-Blackwell.

NTOUTSI E., FAFALIOS P., GADIRAJU U., IOSIFIDIS V., NEJDL W., VIDAL M.-E., RUGGIERI S., TURINI F., PAPADOPOULOS S., KRASANAKIS E., KOMPATSIARIS I., KINDER-KURLANDA K., WAGNER C., KARIMI F., FERNANDEZ M., ALANI H., BERENDT B., KRUEGEL T., HEINZE C., BROELEMANN K., KASNECI G., TIROPANIS T. & STAAB S. (2020). Bias in Data-driven AI Systems - An Introductory Survey. *arXiv preprint 2001.09762*.

OCHS E. (1992). *Indexing gender*, In A. DURANTI & C. GOODWIN, Éds., *Rethinking Context : Language as an interactive phenomenon*, p. 335—350. Cambridge University Press.

PALLETT D. S. (1985). Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards*, **90**, 371–387.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : an ASR corpus based on public domain audio books. In *Actes de ICASSP 2015 (Acoustics, Speech and Signal Processing*, p. 5206–5210, Brisbane, Australie : IEEE.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *Actes de IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society.

SÁNCHEZ-MONEDERO J., DENCIK L. & EDWARDS L. (2020). What does it mean to solve the problem of discrimination in hiring ? social, technical and legal perspectives from the uk on automated hiring systems. In *Actes de FAT 2020 (Fairness, Accountability and Transparency)*, Barcelona, Spain : ACM.

SAON G., KURATA G., SERCU T., AUDHKHASI K., THOMAS S., DIMITRIADIS D., CUI X., RAMABHADRAN B., PICHENY M., LIM L.-L., ROOMI B. & HALL P. (2017). English conversational telephone speech recognition by humans and machines. In *Actes de INTERSPEECH 2017 (International Speech Communication Association)*, p. 132–136 : ISCA. DOI : 10.21437/Interspeech.2017-405.

TATMAN R. (2017). Gender and dialect bias in youtube's automatic captions. In *Actes de ACL Workshop on Ethics in Natural Language Processing*, p. 53–59.

TATMAN R. & KASTEN C. (2017). Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Actes de INTERSPEECH 2017 (International Speech Communication Association)*, p. 934–938 : ISCA.

VANMASSENHOVE E., HARMEIER C. & WAY A. (2018). Getting gender right in neural machine translation. In *Actes de EMNLP 2018 (Empirical Methods in Natural Language Processing)*, p. 3003–3008.

WATANABE S., HORI T., KARITA S., HAYASHI T., NISHITOBA J., UNNO Y., ENRIQUE YALTA SOPLIN N., HEYMANN J., WIESNER M., CHEN N., RENDUCHINTALA A. & OCHIAI T. (2018). Espnet : End-to-end speech processing toolkit. In *Actes de INTERSPEECH 2018 (International Speech Communication Association)*, p. 2207–2211, Hyderabad, India : ISCA. DOI : 10.21437/Interspeech.2018-1456.

XIONG W., WU L., ALLEVA F., DROPPO J., HUANG X. & STOLCKE A. (2018). The Microsoft 2017 conversational speech recognition system. In *Actes de ICASSP 2018 (Acoustics, Speech and Signal Processing*, p. 5934–5938, Calgary, Alberta, Canada : IEEE.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ?

Adam Lion-Bouton[1], Loïc Grobol[2, 3], Jean-Yves Antoine[1], Sylvie Billot[4], Anaïs Lefeuvre-Halftermeyer[4]
(1) LIFAT, ICVL, Université de Tours, 41000 Blois,France
(2) LLF, 8, Rue Albert Einstein 75013 Paris, France
(3) Lattice, 1 Rue Maurice Arnoux, 92120 Montrouge, France
(4) LIFO, ICVL, Université d'Orléans, 45000 Orléans, France
lion.adam.otman@gmail.com, loic.grobol@ens.psl.eu,
Jean-Yves.Antoine@univ-tours.fr,
{Sylvie.Billot, Anais.Halftermeyer}@univ-orleans.fr

RÉSUMÉ

Cet article présente un travail qui consiste à étudier si les scores les plus utilisés pour l'évaluation de la résolution des coréférences constituent des métriques de similarité normalisées. En adoptant une démarche purement expérimentale, nous avons vérifié si les scores MUC, B^3, CEAF, BLANC, LEA et le meta-score CoNLL respectent les bonnes propriétés qui définissent une telle métrique. Notre étude montre que seul le score $CEAF_m$ est potentiellement une métrique de similarité normalisée.

ABSTRACT

Do the standard scores of evaluation of coreference resolution constitute metrics ?

This paper presents an experimental research that investigates whether the most commonly used scores for evaluating the resolution of co-references constitute normalized similarity metrics. Considering systematic test suites, we verified whether the MUC, B3, CEAF, BLANC, LEA and CoNLL scores comply with the formal properties that define such a metric. Our study shows that only the CEAFm score is potentially a normalized similarity metric.

MOTS-CLÉS : coréférence, évaluation, métrique de similarité, MUC, B^3, CEAF, BLANC, LEA, CoNLL.

KEYWORDS: coreference, evaluation, similarity metric, MUC, B^3, CEAF, BLANC, LEA, CoNLL.

1 Introduction

Disposant de ressources linguistiques d'envergure servant aussi bien à l'apprentissage de modèles qu'à leur test, le TAL a recours à des mesures quantitatives pour évaluer ses avancées et étudier la pertinence d'approches alternatives. Généralement, l'évaluation consiste à comparer les réponses du système à une référence idéale (appelée GOLD standard), à l'aide d'un score bien choisi. Ce rôle central de l'évaluation quantitative interroge notre discipline d'un point de vue méthodologique et déontologique. En effet, si une réflexion éthique en recherche doit concerner les productions de cette recherche, elle doit également concerner la manière dont ces recherches sont conduites. On sait en effet que par delà le modèle poppérien de réfutabilité, la recherche est une activité sociale (Latour &

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 10–18.
HAL : hal-02750222.

Woolgar, 1986) dont il est bon d'interroger les pratiques. Lors de la mise en place d'une campagne d'évaluation, de nombreuses questions méritent ainsi d'être posées :

— quelles sont la qualité et la représentativité des données de test utilisées ?
— comment s'assurer de la significativité statistique des différences de performances observées ?
— comment interpréter les résultats de l'évaluation en termes de qualité perçue par l'utilisateur ?
— quels biais peuvent altérer l'interprétation des résultats obtenus ?
— à quelle question répond réellement un score d'évaluation donné, et quel score est le plus appropriée dans une situation donnée ?

Le choix d'un score pertinent est assez évident dans les cas simples. Si l'on considère une tâche de catégorisation telle que l'attribution d'une valence émotionnelle (positif, négatif, aucune) à un tour de parole, le recours à des scores standards pour la classification (rappel, précision, F-mesure) est assez naturel. Notons toutefois que retenir la F-mesure (non pondérée) comme juge de paix pose déjà question : certaines applications peuvent privilégier un besoin en rappel ou au contraire en précision, plutôt que de viser l'optimisation conjointe des deux, telle qu'évaluée par la F-mesure.

Le choix d'un score adéquat est au contraire bien moins évident dans le cas des systèmes end-to-end de TAL, et ce pour deux raisons principales :

— d'une part, il n'est pas toujours évident de définir une référence qui propose une vérité terrain incontestable, de même qu'il n'est pas facile d'évaluer qualitativement si un écart à la référence est plus grave qu'un autre. Si nous prenons ainsi l'exemple de la traduction automatique, comment caractériser une bonne traduction ?
— d'autre part, les tâches complexes ajoutent souvent au problème de la catégorisation celui de la segmentation en unité, ou de l'alignement entre unités. Là encore, il est alors délicat de savoir quelle segmentation est plus contestable qu'une autre

Face à ces difficultés conceptuelles, la communauté scientifique a souvent proposé plusieurs fonctions d'évaluation alternatives pour une tâche donnée. Des études ont été conduites pour étudier les biais et limites de chaque proposition, sans arriver le plus souvent à démêler cet écheveau. Cette situation était prévisible, car souvent, chaque score répond à un objectif d'évaluation particulier. Ce constat se retrouve précisément dans la problématique étudiée dans cet article : la résolution des coréférences. De multiples fonctions d'évaluation ont été proposées, qui répondent chacune à des manières peu conciliables d'appréhender la segmentation d'un texte en chaînes de références. Ceci conduit à des résultats sensiblement différents suivant le score utilisée (Moosavi & Strube, 2016)

Le travail présenté dans cet article vise à clarifier cette situation en vérifiant quelles propriétés formelles vérifient les scores les plus utilisés. Plus précisément, nous proposons de vérifier quels scores peuvent être considérés comme des métriques de similarité normalisées. Ces contraintes formelles ne constituent pas le seul critère de choix d'un score d'évaluation. Elles ne rendent pas compte des biais statistiques et ne parlent que de manière indirecte des liens qui existent entre évaluation objective et jugement humain. Si l'on cherche à mesurer la qualité d'un système, il nous semble toutefois important d'étudier les propriétés formelles de l'outil de mesure utilisé.

Dans un premier temps, nous présentons la tâche de résolution des coréférences, puis les différents scores qui ont été proposés pour l'évaluer. La section suivante décrit les propriétés formelles qui définissent une métrique de similarité normalisée. Nous présentons ensuite notre protocole d'étude. La présentation des résultats est enfin l'occasion de donner un tableau synthétique de l'intérêt de ces scores d'évaluation très répandus.

2 L'évaluation pour la coréférence

On définit une **mention** comme un mot ou groupe de mots référant à un élément de l'univers du discours appelé **entité**, et une **chaîne de coréférences** comme l'ensemble des mentions référant à une même entité. On dit de deux mentions référant à une même entité qu'elles sont **coréférentes** et on nomme **singleton** toute entité qui ne compte qu'une unique mention. Par définition, une mention réfère nécessairement à une unique entité. L'ensemble des chaînes de coréférences d'un document forme donc une partition de l'ensemble des mentions de ce document, dont les chaînes de coréférences sont les parties. Considérons l'énoncé suivant :

> Elles$_1$ tournent, tournent les aiguilles$_1$, ne m'$_2$ dis plus qu'c'est passager
>
> Furax Barbarossa - Fin 2006

Dans cet exemple ainsi que dans les suivants, chaque mention sera indicée par le numéro de l'entité auquel elle réfère. Cet exemple est composé de trois mentions et des deux entités suivantes :

1 : Elles , les aiguilles
2 : m' (singleton)

Étant donné un document et l'ensemble des mentions le composant [1], la tâche de résolution des chaînes de coréférences consiste à produire une partition de mentions (que l'on appellera SYS) aussi proche de la vérité (appelé GOLD) que faire se peut. L'évaluation consiste ainsi à attribuer un score à la partition SYS en fonction de sa similarité avec la partition GOLD. Or, il s'avère qu'une partition peut être caractérisée de plusieurs manières, induisant chacune différentes méthodes de comparaison. Deux mentions coréférentes peuvent par exemple être vues respectivement comme appartenant à un même ensemble, référant à une même entité, ou encore comme étant **liées** par un *lien* (relation) de coréférence. Il existe alors plusieurs manières d'évaluer les différences (et donc les similarités) entre le GOLD et le SYS suivant la vision de la coréférence adoptée.

Considérons par exemple les deux partitions suivantes — respectivement GOLD et SYS :

> J'$_1$ suis bien chez moi$_1$ mais j'$_1$ comprends l'gitan$_2$ quand il$_2$ m'$_1$ rappelle qu' il$_2$ est libre
> J'$_1$ suis bien chez moi$_1$ mais j'$_3$ comprends l'gitan$_2$ quand il$_2$ m'$_3$ rappelle qu' il$_2$ est libre
>
> Furax Barbarrossa - Fin 2012

Une première approche représente une chaîne de coréférence comme la suite des liens entre les mentions qui la composent, en respectant leur *succession* dans le texte. Ainsi, sur la figure 1.a, on décrit l'entité 1 du GOLD par les trois liens J'$_1$ — moi$_1$, moi$_1$ — j'$_1$, j'$_1$ — m'$_1$, tandis que les entités 1 et 3 détectées dans SYS se décrivent par les deux liens J'$_1$ — moi$_1$, j'$_3$ — m'$_3$. Dans cette approche, ajouter au SYS le lien moi — j' fusionnerait les entités 1 et 3 du SYS, résultant en un SYS égal au GOLD. En somme, GOLD ne diffère de SYS que par une seule opération.

On pourrait au contraire considérer qu'une chaine de coréférences est décrite par l'ensemble de *tous* les liens possibles entre les mentions qui la composent (figure 1.b), indépendamment de leur ordre d'occurrence. Dans cette approche, l'entité 1 du GOLD est décrite par les six liens J'$_1$ — moi'$_1$, J'$_1$ — j'$_1$, J'$_1$ — m'$_1$, moi$_1$ — j'$_1$, moi$_1$ — m'$_1$, j'$_1$ — m'$_1$ et les entités 1 et 3 du SYS

1. On ne s'intéresse ici pas à la tâche consistant à retrouver les mentions. De même, notre étude ne concerne que la résolution des coréférences, et non celle des anaphores pour lesquelles l'identification d'une référence est nécessaire pour l'interprétation d'une mention, sans qu'il y ait nécessairement identité de référence

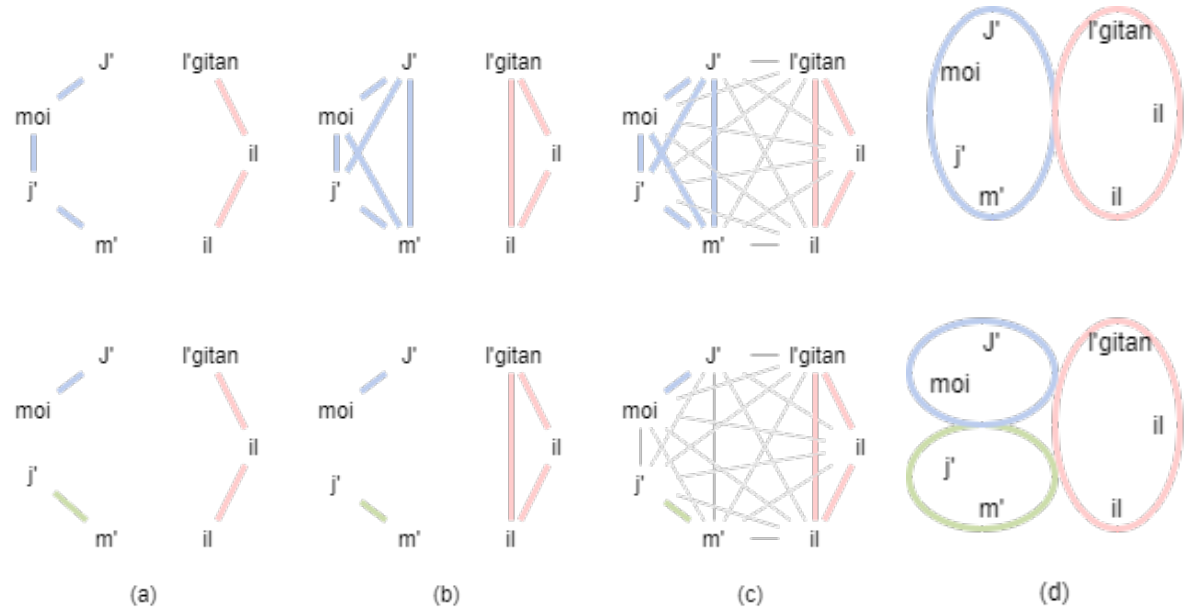


FIGURE 1 – Représentations des partitions de nos exemple GOLD (en haut) et SYS (en bas) suivant différentes approches de la coréférence : (a) chaînes de maillons (b) ensemble de liens de coréférence (c) ensemble de liens de coréférence, en gras et de non coréférence, en traits fin (d) vision ensembliste

par les deux liens J'$_1$ — moi$_1$, j'$_3$ — m'$_3$. Dans ce cas, il faudrait ajouter au SYS les quatre liens suivants pour retrouver le GOLD : J' — j', J' — m', moi — j', moi — m' pour que SYS et GOLD soient égaux.

Ce raisonnement peut être poursuivi dans une approche qui se focalise sur une vision ensembliste (figure 1 ;d). Dans ce cas, il faut déplacer les mentions j' et m' dans l'ensemble représentant l'entité 1 pour identifier GOLD et SYS. Soit donc deux opérations. Arrêtons là la démonstration pour simplement constater que plusieurs scores ont été proposés, qui répondent à ces différentes approches :

— **MUC**, proposé par Vilain *et al.* (1995) comme fonction d'évaluation de la tâche MUC6 (MUC Consortium, 1995). MUC considère le nombre minimum de liens nécessaires pour décrire les deux partitions et calcule une similarité en fonction de la proportion de ces liens correctement résolus (i.e. qui sont communs au GOLD et au SYS).
— **B^3** proposé par Bagga & Baldwin (1998) s'intéresse, pour toutes les paires d'entités de GOLD et de SYS, au nombre de mentions correctement résolues.
— **CEAF**, proposé par Luo (2005), choisit le meilleur appariement entre les entités GOLD et SYS puis s'intéresse à la proportion de mentions correctement résolues, mais uniquement pour les entités ainsi appariées. La définition de meilleur appariement n'a pas une solution unique, Luo (2005) propose ainsi deux alternatives, l'une centrée sur les mentions, nommée **CEAF$_\mathbf{m}$**, l'autre centrée sur les entités et nommée **CEAF$_\mathbf{e}$**
— **BLANC** (Recasens & Hovy, 2011) considère les mentions non-coréférentes comme liées par un lien de non-coréférence (figure 1.c) et évalue la proportion de liens (de coréférences *et* de non-coréférence) correctement résolus en s'abstrayant de la notion d'entité.
— **LEA** (Moosavi & Strube, 2016) s'intéresse à la fraction de liens de coréférences bien résolus.

Aucun de ces scores n'est parvenu à faire l'unanimité. Face à cela, Denis & Baldridge (2009) proposent

MELA, une moyenne de MUC, B^3, et $CEAF_e$, employé par Pradhan *et al.* (2012) comme fonction d'évaluation de la tâche **CoNLL**-2012 (et pour cette raison souvent appelée «score CoNLL»).

Ces fonctions de scores sont définies par une évaluation de la proximité entre SYS et une unique partition GOLD. Dans une perspective d'interprétabilité, il est intéressant de les étudier plus généralement comme des estimations de la similarité entre deux partitions, sans nécessairement que l'une d'entre elles ait un rôle privilégié. Ainsi, il serait par exemple souhaitable que deux sorties systèmes proches donnent également des scores proches quand elles sont comparées au GOLD et, par extension, que deux systèmes donnant des sorties similaires obtiennent également des scores proches.

Dans une perspective de conception de systèmes, pouvoir estimer directement la proximité de sorties systèmes sans passer par une référence serait également souhaitable, par exemple pour étudier la sensibilité d'un système au variations dans un choix de paramètres. Dans la suite de cet article, nous étudions précisément les propriétés formelles de ces scores, non pas en termes d'adéquation avec une référence mais en terme de similarité entre deux partitions. Pour ce faire, nous nous intéressons en particulier aux propriétés définissant une métrique de similarité normalisée.

3 Métrique de similarité normalisée

La notion de métrique de distance fait l'objet d'une définition reconnue, à l'inverse de celle de similarité qui est le plus souvent ramenée au complément d'une distance (Deza & Deza, 2009). Pour répondre à cette carence formelle, Chen *et al.* (2009) dérive de la définition d'une métrique de distance celle d'une métrique de similarité : on appelle **métrique de similarité** toute fonction $s : X \times X \longrightarrow \mathbb{R}$ respectant les cinq propriétés suivantes pour tout $(a, b, c) \in X^3$:

$$s(a,b) = s(b,a) \quad \text{(symétrie)} \tag{1}$$
$$s(a,a) \geq 0 \tag{2}$$
$$s(a,b) \leq s(a,a) \tag{3}$$
$$s(a,b) + s(b,c) \leq s(b,b) + s(a,c) \quad \text{(inégalité triangulaire)} \tag{4}$$
$$s(a,a) = s(b,b) = s(a,b) \text{ si et seulement si } a = b \quad \text{(identité des indiscernables)} \tag{5}$$

On dira de plus que s est **normalisée** si

$$s(a,b) \leq 1 \tag{6}$$

Enfin, si s est une métrique de similarité normalisée telle que

$$s(a,a) = 1 \quad \text{(identité à 1)} \tag{7}$$
$$s(a,b) \geq 0 \quad \text{(positivité)} \tag{8}$$

alors $1 - s$ est une métrique de *distance* normalisée. Dans ce cas, les propriétés 2 à 5 peuvent être réécrites sous une forme plus naturelle :

$$s(a,a) = 1 \quad \text{(équivalent à la propriété 7)} \tag{9}$$
$$s(a,b) \leq 1 \quad \text{(équivalent à la propriété 6)} \tag{10}$$
$$s(a,b) + s(b,c) \leq 1 + s(a,c) \tag{11}$$
$$s(a,b) = 1 \Leftrightarrow a = b \tag{12}$$

Vérifier si un score est une métrique de similarité normalisée nous parait important. En effet, une métrique de similarité normalisée dispose de bonnes propriétés qui rendent intelligibles les écarts de performance qu'elle peut mesurer. Lorsque l'on compare les performances de différents systèmes, on opère des jugements cognitifs de similarité. À la suite de Shepard (1962), de multiples travaux en psychologie ont suggéré que ces jugements relèvent d'une évaluation de distance dans un espace métrique, ceci même si certains biais peuvent conduire à des violations de la contrainte de métricité sur des cas limites (Laub *et al.*, 2007). Ces violations n'empêchent pas de même Tversky & Gati (1982) de considérer que la propriété d'inégalité triangulaire favorise l'intelligibilité des comparaisons.

D'un point de vue pratique, le statut de métrique de similarité est préservé par plusieurs opérations élémentaires telles que la somme et le produit avec une autre métrique de similarité, ainsi que par l'application d'une fonction convexe positive strictement croissante (Chen *et al.*, 2009). Cette dernière opération permet un rééquilibrage du score pour obtenir une distribution plus uniforme des valeurs sur l'intervalle $[0, 1]$ et donc d'obtenir un score plus aisément interprétable, tandis que somme et produit sont utiles pour construire des scores d'évaluation combinés tel que CoNLL.

4 Protocole

Pour cette étude, nous avons défini un protocole de vérification des propriétés précédentes pour chacun des scores présentés plus haut. On adopte une démarche qui a pour objectif non pas de valider ces propriétés (ce qui ne peut se faire que formellement), mais de tenter de les falsifier à l'aide de jeux de tests bien choisis . Cette méthode a l'avantage de s'appliquer à toute fonction de score, que son fonctionnement interne soit connu ou pas. Elle s'applique par ailleurs à des fonctions de score qui ne seraient pas définies formellement, mais répondraient par exemple à des estimations statistiques. Notons que dans les cas de falsification, notre démarche a force de preuve analytique : nous sommes en effet en mesure de donner un contre-exemple.

La démarche employée tire parti de deux traits spécifiques à l'évaluation de la coréférence. Premièrement, les scores ne sont calculés que pour évaluer des partitions d'un même ensemble de mentions. Deuxièmement, la similarité entre deux partitions ne dépend que des relations définies par les partitions, et aucunement d'une éventuelle sémantique propre aux mentions. Ainsi, une propriété vérifiée pour toutes les partitions d'un ensemble de n éléments particuliers le sera également pour toutes les partitions de *tous* les ensembles à n éléments. Réciproquement, un contre-exemple valable pour un ensemble à n éléments particulier le sera pour tous les ensembles à n éléments.

$$s(\{\{1,2\}\},\{\{1,2\}\}) \overset{?}{\leq} s(\{\{1,2\}\},\{\{1,2\}\})$$
$$s(\{\{1,2\}\},\{\{1\},\{2\}\}) \overset{?}{\leq} s(\{\{1,2\}\},\{\{1,2\}\})$$
$$s(\{\{1\},\{2\}\},\{\{1,2\}\}) \overset{?}{\leq} s(\{\{1\},\{2\}\},\{\{1\},\{2\}\})$$
$$s(\{\{1\},\{2\}\},\{\{1\},\{2\}\}) \overset{?}{\leq} s(\{\{1\},\{2\}\},\{\{1\},\{2\}\})$$

FIGURE 2 – Exemple de la procédure de falsification

Suivant ce principe, pour chaque propriété — par exemple : (3) $(s(a,b) < s(a,a)$ — on génère toutes les partitions de l'ensemble des n premiers nombres entiers positifs, et on remplace chacune des variables dans la définition de la propriété — a et b pour la propriété 3 — par toutes les combinaisons de partitions de cet ensemble. Ainsi, si l'on considère $n = 2$, on obtient deux partitions : $\{\{1,2\}\}$ et $\{\{1\},\{2\}\}$. On évalue alors les cas présentés en figure 2.

On procède exhaustivement jusqu'à $n = 5$ dans le cas de la propriété 4 (soit environs 140 000 combinaisons à évaluer) et jusqu'à $n = 6$ pour les autres (environ 40 000 combinaisons par propriété).

5 Résultats

Les résultats des expérimentations sont résumés dans la table 1.

	1	2	3	4	5	6	7	8
MUC				X			X	
B^3				X				
$CEAF_m$								
$CEAF_e$				X				
CoNLL				X			X	
BLANC	X			X				
LEA				X				

TABLE 1 – Propriétés non-respectées par les scores : (X) si la propriété n'est pas respectée

Symétrie — On remarque tout d'abord que BLANC ne respecte pas la propriété 1 et n'est donc pas symétrique. Par exemple, on observe que :

$$BLANC(\{\{1,2,3\}\},\{\{1,2\},\{3\}\}) = 0{,}5 \neq 0{,}25 = BLANC(\{\{1,2\},\{3\}\},\{\{1,2,3\}\})$$

Nos tests montrent que BLANC est le plus souvent symétrique. La propriété n'est enfreinte que lorsqu'une des partitions considérées est composée uniquement de singletons ou d'une entité unique. BLANC est en effet défini par une formule générale, mais aussi par des cas particuliers dans ces situations. En l'occurrence, ces cas particuliers s'avèrent être la cause de la non-symétrie du score.

On pourrait penser que ces cas particuliers sont rares et peuvent être négligés. Il convient cependant de se rappeler que plus l'ensemble de mentions considéré est petit, plus la probabilité de rencontrer l'un de ces cas particuliers est élevée. Ainsi, 6 des 15 combinaisons de deux partitions créées sur un ensemble de trois mentions débouchent sur l'un de ces cas particuliers non-symétriques.

Identité à 1 — En second lieu, on remarque que ni MUC, ni CoNLL ne respectent la propriété 7, d'identité à 1 lorsqu'une partition est comparée à elle-même. Comme précédemment, cette propriété semble être généralement respectée et uniquement enfreinte dans le cas où la partition comparée à elle-même est composée uniquement de singletons.

MUC a été définie dans le cadre d'une tâche où la notion de singleton était ignorée. MUC n'a donc pas été conçue pour en tenir compte et vaut 0 dans le cas où l'une des partitions n'est composée que de singletons. CoNLL étant défini comme une moyenne de scores intégrant MUC, lorsque MUC est égal à 0, CoNLL ne peut nécessairement pas être supérieur à $\frac{2}{3}$.

Un score ne respectant pas l'identité à 1 a pour défaut de rendre l'interprétation de score plus complexe. Dans le cas présent, afin d'être correctement interprétés les scores MUC et CoNLL devraient être accompagnés d'une information supplémentaire, indiquant si l'une des partitions est composée uniquement de singleton.

Inégalté triangulaire — Enfin, on remarque que ni MUC, B^3, $CEAF_e$, CoNLL, BLANC, ni LEA ne respectent la propriété 4, aucun de ces scores ne respecte donc l'inégalité triangulaire. Contrairement à précédemment, cette propriété est falsifiée même en dehors de cas limites. Les trois exemples suivants falsifient respectivement l'inégalité triangulaire pour les scores MUC, B^3 et LEA ; les scores $CEAF_e$ et BLANC ; et le score CoNLL.

$$a = \{\{1,2,3\},\{4,5\}\}, b = \{\{1,2,3\},\{4\},\{5\}\}, c = \{\{1,2\},\{3\},\{4\},\{5\}\}$$
$$a = \{\{1,2,3,4\},\{5\}\}, b = \{\{1,2,3\},\{4\},\{5\}\}, c = \{\{1,2,3,5\},\{4\}\}$$
$$a = \{\{1,2,3,4\},\{5\}\}, b = \{\{1,2,3\},\{4,5\}\}, c = \{\{1\},\{2\},\{3\},\{4,5\}\}$$

Un score ne respectant pas cette propriété pose de forts problèmes d'interprétation. Il est alors possible pour deux partitions SYS très similaires d'obtenir des scores très différents. L'inégalité triangulaire est ainsi la propriété donnant son coeur à la notion de métrique.

6 Conclusion

Dans cet article, nous nous sommes interrogés sur les propriétés formelles que vérifient les principaux scores utilisés pour évaluer la résolution de la coréférence. Nous avons vu que ces scores diffèrent par la vision qu'ils proposent de la coréférence Afin de comparer ces scores, nous avons considéré les bonnes propriétés qu'ils doivent vérifier afin d'être considérés comme des métriques de similarité normalisées. Nous avons alors cherché à falsifier par des jeux de test l'hypothèse selon laquelle les scores étudiés constituent (ou non) de telles métriques. Cette étude a prouvé que MUC, B^3, $CEAF_e$, CoNLL, BLANC et LEA ne sont pas des métriques alors que $CEAF_m$ semble être une métrique de similarité normalisée. Du moins, il ne nous a pas été possible de falsifier cette hypothèse.

$CEAF_m$ apparaît ainsi comme une métriques intéressante d'un point de vue formel, car son possible statut de métrique de similarité normalisée répond aux modèles cognitifs décrivant les processus de catégorisation, de comparaison et d'estimation de ressemblances réalisés par les humains. Par delà cette étude purement formelle, il est toutefois important de s'intéresser également aux biais statistiques (influence de la taille des entités, influence de la prévalence des singletons, etc.) qui peuvent influencer les scores d'évaluation obtenus par ces métriques. Nous sommes précisément en train d'étudier systématiquement ces types de biais, à l'aide d'une approche expérimentale sur jeux de tests. Nous étudions également le degré de corrélation qui existe entre scores d'évaluation et jugements humains de la qualité des systèmes (Holen, 2013). Ce tableau complet (propriétés formelles, tolérance aux biais, proximité avec une évaluation subjective) devrait alors permettre au concepteur de système ou de campagne d'évaluation de choisir, en toute connaissance, le score d'évaluation le plus adapté à ses besoins, ou à mieux connaître les avantages et faiblesses de chacun.

Remerciements

Ce travail a bénéficié du soutien du RTR DIAMS de la région Centre-Val-de-Loire.

Références

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, p. 563–566.

CHEN S., MA B. & ZHANG K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science*, **410**(24-25), 2365–2376.

DENIS P. & BALDRIDGE J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, **42**.

DEZA M. M. & DEZA E. (2009). Encyclopedia of distances. In *Encyclopedia of distances*, p. 1–583. Springer.

HOLEN G. I. (2013). Critical reflections on evaluation practices in coreference resolution. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, p. 1–7.

LATOUR B. & WOOLGAR S. (1986). *Laboratory Life : The Construction of Scientific Facts*, volume 80. Princeton University Press.

LAUB J., MÜLLER K.-R., WICHMANN F. A. & MACKE J. H. (2007). Inducing metric violations in human similarity judgements. In *Advances in neural information processing systems*, p. 777–784.

LUO X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, p. 25–32 : Association for Computational Linguistics.

MOOSAVI N. S. & STRUBE M. (2016). Which coreference evaluation metric do you trust ? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 632–642.

MUC CONSORTIUM (1995). Appendix D : Coreference Task Definition (v2.3). In *Sixth Message Understanding Conference*, Columbia, Maryland : Morgan Kaufmann.

PRADHAN S., MOSCHITTI A., XUE N., URYUPINA O. & ZHANG Y. (2012). Conll-2012 shared task : Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, p. 1–40 : Association for Computational Linguistics.

RECASENS M. & HOVY E. (2011). Blanc : Implementing the rand index for coreference evaluation. *Natural Language Engineering*, **17**(4), 485–510.

SHEPARD R. N. (1962). The analysis of proximities : multidimensional scaling with an unknown distance function. i. *Psychometrika*, **27**(2), 125–140.

TVERSKY A. & GATI I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, **89**(2), 123.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, p. 45–52 : Association for Computational Linguistics.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Que recèlent les données textuelles issues du web ?

Adrien Barbaresi[1] Gaël Lejeune[2]
(1) Académie des Sciences de Berlin-Brandenbourg, Jägerstraße 22-23, 10117 Berlin, Allemagne
(2) Sorbonne Université, 1 rue Victor Cousin, 75005 Paris, France

RÉSUMÉ

La collecte et l'usage opportunistes de données textuelles tirées du web sont sujets à une série de problèmes éthiques, méthodologiques et épistémologiques qui méritent l'attention de la communauté scientifique. Nous présentons des études empiriques de leur impact en linguistique et TAL centrées sur la forme (méthodes d'extraction des données) ainsi que sur le fond (contenu des corpus).

ABSTRACT

What do text data from the Web have to hide ?

The opportunistic gathering and use of text data taken from the Web are subject to a whole series of ethical, methodological and epistemological problems which could benefit from the interest of the research community. We present empirical studies of their impact in linguistics and natural language processing, with respect to their form (extraction methods) and to their contents.

MOTS-CLÉS : Construction de corpus, Science du web, Extraction de texte, Méthodes d'évaluation.

KEYWORDS: Corpus construction, Web science, Text extraction, evaluation methods.

1 Introduction

Le web est fréquemment perçu comme un « réservoir indifférencié de textes à analyser » pour le TAL (Tanguy, 2013) et l'on peut affirmer sans risque que web et TAL poursuivent leur « histoire commune » (*op. cit.*). Les données issues du web y sont en effet omniprésentes, à la fois en tant qu'instantané d'un état de la langue, de données à analyser pour elles-mêmes, mais aussi de références destinées à construire des modèles de langue ou des ressources langagières. De la collecte au corpus, non seulement opportuniste (McEnery & Hardie, 2011) mais également « prêt-à-utiliser », il n'y a souvent qu'un pas. Le Common Crawl [1] notamment s'est imposé comme source majeure pour des tâches variées, de la traduction automatique neurale (Smith *et al.*, 2013) à la construction et (dans une situation optimale) à l'affinage de modèles de langue basés sur des techniques d'apprentissage profond nécessitant des données massives (Suárez *et al.*, 2019). Cette évolution a conduit à des problèmes récurrents d'ordre éthique, à l'image du robot conversationnel Tay lancé par Microsoft en 2016 sur Twitter et stoppé 16 heures après son entrée en fonction en raison de l'ampleur et de la gravité des messages racistes et sexistes « appris » et ensuite (re-)publiés par le robot [2]. De même, des modèles entraînés sur des données massives (d'origine contrôlée ou non) intègrent une série de biais sociétaux (Caliskan *et al.*, 2017). Malgré une certaine impression de facilité quant à la construction de corpus, les méthodes utilisant des corpus web nécessitent des dispositifs expérimentaux et des

1. https ://commoncrawl.org
2. https ://en.wikipedia.org/wiki/Tay_(bot)

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 19–28.
HAL : hal-02750223.

instruments ad hoc (Valette, 2008) afin d'estimer leur qualité et leur adéquation aux tâches proposées. D'un point de vue épistémologique, la simple accumulation de données textuelles ne rend pas pour autant ce terrain intelligible et un retour analytique sur ces données s'avère nécessaire [3].

L'objet de cet article est de documenter et de commenter des problèmes liés au contenu des corpus web ainsi qu'aux processus d'extraction qui permettent d'accéder aux textes. Nous considérons des corpus construits directement sans recours à des données pré-existantes et utilisons à cette fin des méthodes de parcours du web, *web crawling* (Olston & Najork, 2010), afin de découvrir des hôtes hébergeant des pages web d'une part (méthode « généraliste ») et des documents au sein de domaines déjà connus (méthode « ad hoc ») d'autre part (Barbaresi, 2015). La simple notion de texte comme unité langagière cohérente est sujette à caution sur ces données reprises et potentiellement déformées par les outils de traitement. Nous souhaitons apporter un retour quantitatif, avec un examen des méthodes d'extraction, ainsi qu'un examen qualitatif, avec l'exemple des discours haineux, afin de remettre en question l'usage sans garde-fous des textes tirés du web et des informations qu'ils recèlent.

2 Examen des méthodes d'extraction

Étant donné le code source d'une page web, le processus d'extraction de contenu consiste à détourer le contenu textuel utile (c'est-à-dire notamment sans les éléments de structure, la publicité ou encore les commentaires) et à identifier les méta-données. Concrètement, cette tâche implique une conversion du format HTML vers un autre format (souvent plein texte ou XML). La construction de corpus à partir du web est devenue un élément si commun des chaînes de traitement de TAL que les détails techniques sur sa mise en œuvre sont souvent omis. Or, peut-on réellement s'abstenir de se demander ce qui est intégré dans des corpus ? Afin d'illustrer ce problème de contenu, nous comparons différents extracteurs récents et/ou populaires afin d'observer ce que des métriques d'évaluation spécialement conçues révèlent concernant leur efficacité. Nous laisserons de côté le choix des sources en elles-mêmes pour nous concentrer sur les résultats de l'extraction, qui forment la base de décisions quant à l'inclusion dans le corpus final (Schäfer *et al.*, 2013), et nous intéressons en particulier à la question multilingue. En effet, les outils mis à disposition de la communauté sont très souvent conçus pour la langue anglaise, l'applicabilité à d'autres langues étant souvent considérée de manière opportuniste comme allant de soi : si les extracteurs fonctionnent sur l'anglais, alors les mêmes ordres de grandeur de résultats seront obtenus dans d'autres langues.

Corpus et outils Nous reprenons un corpus proposé par Lejeune & Zhu (2018) qui comprend près de 1.700 documents en 5 langues (475 en anglais, 405 en chinois, 273 en grec, 274 en polonais et 267 en russe) avec la version HTML d'une part et une version de référence nettoyée manuellement d'autre part. La sélection des outils s'est faite sur trois critères : simplicité d'utilisation (existence d'une version ou *wrapper* PYTHON puisque ce langage est très répandu, en particulier en TAL) ; disponibilité sous licence libre (pour les mêmes raisons, nous tenons uniquement compte d'outils directement accessibles) ; popularité ou nouveauté (l'état de l'art fournit des informations, notamment pour des outils relativement anciens comme BOILERPIPE).

3. « La collecte et la mise en circulation des données dans des dispositifs adéquats aboutissent à une mise en ordre du monde qui relève d'une cosmétique : ces données sont triées, classées, archivées. Ces opérations textuelles permettent l'accumulation et donc l'archivage mais ne rendent pas pour autant le terrain intelligible. Cette intelligibilité du terrain est le résultat d'une deuxième opération, celle de mise en ordre des données accumulées, de leur traitement, de leur analyse et de leur restitution. » (Calberac, 2010, p. 104)

En conséquence, nous choisissons de comparer les outils suivants, classés en différentes catégories selon leur finalité : conversion du format HTML vers le format texte (I), intégration dans un contexte plus large d'extraction et d'analyse d'information (II), extraction proprement dite du texte principal d'une page web (*boilerplate removal*, III).

Cat.	Outil	Version	Adresse Github	Référence
I	HTML2TEXT	2020.1.16	Alir3z4/html2text	
I	INSCRIPTIS	1.0	weblyzard/inscriptis	
II	NEWSPAPER3K	0.2.8	codelucas/newspaper	
II	NEWS-PLEASE	1.4.25	fhamborg/news-please	(Hamborg *et al.*, 2017)
II	READABILITY	0.7.1	buriy/python-readability	
III	BOILERPY3	1.0.2	jmriebold/BoilerPy3	(Kohlschütter *et al.*, 2010)
III	DRAGNET	2.0.4	dragnet-org/dragnet	(Peters & Lecocq, 2013)
III	GOOSE3	3.1.6	goose3/goose3	
III	JUSTEXT	2.2.0	miso-belica/jusText	(Pomikálek, 2011)
III	TRAFILATURA	0.4.1	adbar/trafilatura	(Barbaresi, 2019)

Ce comparatif fait également l'objet d'une démonstration (Lejeune & Barbaresi, 2020), ces résultats peuvent être reproduits en utilisant les données et scripts mis à disposition [4]. Nous optons ici pour une version abrégée : une seule des configurations de BOILERPIPE (BP3_Article), configuration par défaut pour JUSTEXT et TRAFILATURA.

Mesures Les mesures d'évaluation de la campagne Cleaneval (Baroni *et al.*, 2008) sont fondées sur la préservation des séquences de tokens. Bien qu'imparfaites, elles ont le mérite d'être globalement utilisées par la communauté scientifique (Weninger *et al.*, 2016). Nous ajoutons une mesure plus simple, fondée sur la préservation du vocabulaire, qui donne des résultats tout à fait comparables. Cette évaluation nécessite une vérité de terrain, que nous appellerons GT et GT_{tok} pour la séquence de tokens correspondante. Nous nommons RES le résultat de l'extraction automatique et RES_{tok} la séquence de tokens correspondante, en reprenant le tokeniseur fourni par Cleaneval. La mesure Cleaneval vérifie à quel point la séquence de tokens extraite automatiquement (RES_{tok}) est similaire à la séquence de référence (GT_{tok}). L'algorithme de Ratcliff/Obershelp (Ratcliff & Metzener, 1988) est utilisé pour détecter les plus longues séquences de tokens communes et non-redondantes, sa complexité quadratique est peu efficace et ses résultats ne sont pas immédiatement interprétables. Notre mesure plus simple (*occ_eval*) vérifie si le nombre d'occurrences des tokens correspond aux nombre d'occurrences attendues.

Quelques résultats Nous détaillons ici quelques résultats tirés d'une analyse des outils (Barbaresi & Lejeune, 2020) et nous concentrons sur leur variabilité. Le tableau 1a présente les résultats globaux avec la métrique `clean_eval`. La précision et le rappel sont des moyennes des précisions et rappels par document. La F-mesure est calculée à partir de ces moyennes [5]. Le tableau 1b présente les résultats obtenus avec `occ_eval`, ceux-ci diffèrent assez peu. L'ordre de grandeur des résultats et la « hiérarchie » entre les outils sont conservés à ceci près que JUSTEXT semble pénalisé par la mesure `clean_eval`.

Variation selon les langues D'un point de vue général, l'outil le plus fiable semble être BP3_ART, READABILITY TRAFILATURA et JUSTEXT se situant juste derrière. Toutefois, les moyennes (micro ou macro) masquent des différences entre les langues, comme nous le montrons dans les tableaux 2a

4. https ://www.github.com/run-dimeco/waddle
5. La moyenne des f-mesures donne un score peu intuitif car souvent inférieur à la macro-précision et au macro-rappel.

Outil	Macro F	Micro F	Micro P	Micro R
BP3_Art	72,73	**78,84**	82,80	75,24
READ	74,62	75,87	72,18	79,96
TRAF	**75,69**	75.71	68.33	84.87
JT	63,7	71,22	78,93	64,88
DRAGNET	58,21	69,66	**87,53**	57,85
NPLEASE	48,84	58,46	69,00	50,72
GOOSE	37,87	53,93	83,89	39,74
NPAPER	32,37	50,83	82,20	36,78
INSCRI	40,10	42,95	27,72	**95,28**
HTML2T	31,2	33,98	20,86	91,47

(a) Mesure `clean_eval`

Outil	Macro F	Micro F	Micro P	Micro R
BP3_Art	70,41	**76,38**	80,60	72,57
JT	67,7	74,13	81,36	68,08
READ	71,01	73,25	72,43	74,09
TRAF	68,63	72.89	65.02	82.93
DRAGNET	56,12	67,09	86,82	54,67
NPLEASE	50,92	66,64	**92,03**	52,23
GOOSE	41,72	57,74	89,42	42,64
NPAPER	36,18	54,78	88,68	39,63
INSCRI	34,98	37,10	23,22	**92,22**
HTML2T	30,95	33,45	20,56	89,80

(b) Mesure `occ_eval`

TABLE 1: Evaluation sur le corpus multilingue, F-mesure calculé à partir des micro-moyennes de la Précision et du Rappel (sur fond grisé la différence d'ordre entre les deux mesures)

Outil	F-mes.	Préc.	Rap.
NPAPER	**91,32**	91,34	91,31
GOOSE	90,69	**92,94**	88,54
NPLEASE	88,91	87,89	89,96
DRAGNET	88,78	88,52	89,04
READ	87,16	84,31	90,21
BP3_Art	87,00	87,50	86,51
JT	84,86	83,16	86,62
TRAF	82.58	74.28	92.97
INSCRI	45,84	29,88	**98,46**
HTML2T	44,61	28,98	96,84

(a) `occ_eval` (Anglais)

Outil	F-mes.	Préc.	Rap.
JT	**76.29**	71.64	81.59
READ	74.27	72.29	76.36
TRAF	71.20	64.80	79.02
BP3_Art	69.31	70.11	68.53
DRAGNET	50.94	85.13	36.34
NPLEASE	42.64	93.16	27.64
GOOSE	40.24	90.96	25.83
INSCRI	32.53	19.77	**91.75**
HTML2T	29.55	17.63	91.35
NPAPER	5.14	**92.34**	2.64

(b) `occ_eval` (Russe)

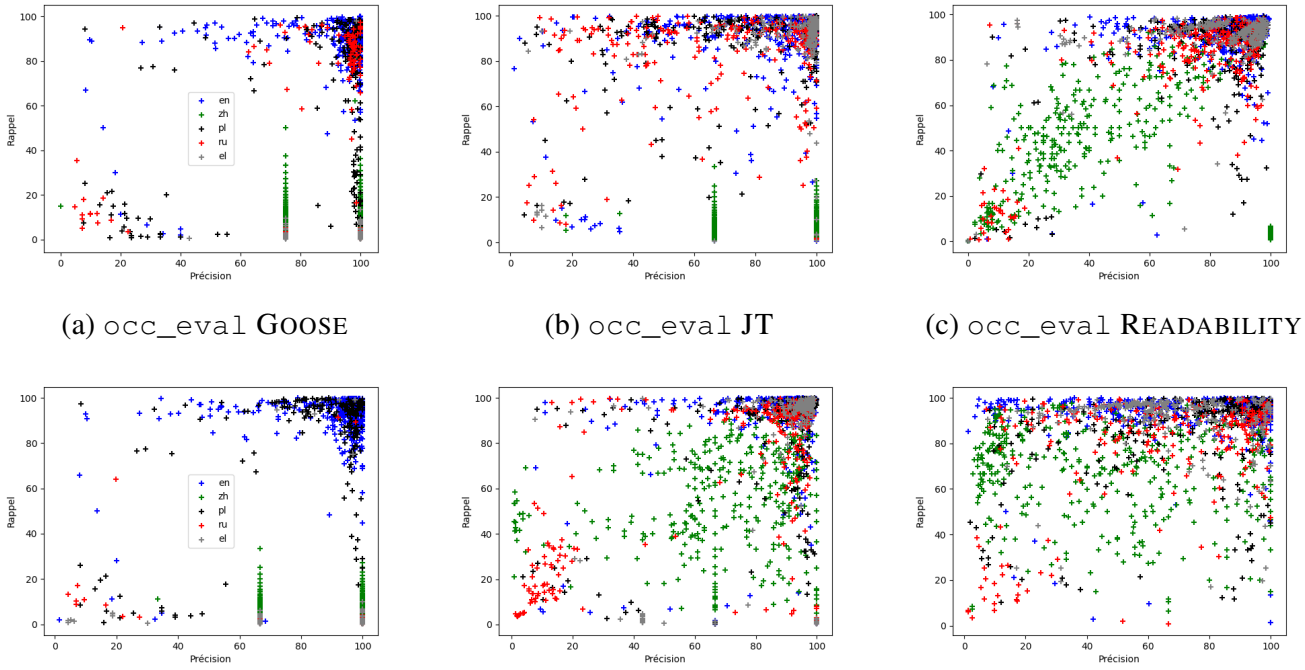
Outil	F-mes.	Préc.	Rap.
BP3_Art	**63.30**	71.28	56.93
TRAF	55.48	46.81	68.09
DRAGNET	44.53	81.81	30.59
READ	42.36	48.00	37.91
GOOSE	20.60	82.54	11.77
JT	19.19	82.32	10.86
NPAPER	19.17	82.72	10.84
HTML2T	13.83	7.62	74.87
NPLEASE	13.31	**97.52**	7.14
INSCRI	12.97	7.06	**79.52**

(c) `occ_eval` (Chinois)

TABLE 2: `occ_eval` par langue, sur fond gris les systèmes les plus performants au global

à 2c qui présentent les résultats sur les sous-corpus anglais, russe et chinois pour une sélection d'outils parmi les plus efficaces. Nous avons marqué en grisé les performances des 4 outils les plus efficaces sur le corpus multilingue, ce qui permet de voir qu'ils sont bien placés, sauf sur le sous-corpus anglais où des outils très spécialisés sont plus performants. L'anglais est évidemment la langue la mieux traitée puisque 9 des 11 systèmes testés ont une F-mesure au dessus de 80% (2 en Grec et 3 en polonais). En ce qui concerne les performances par outil, BP3_ART, le meilleur outil selon la micro-moyenne générale, est inégal selon la langue traitée : très efficace comparativement aux autres sur le chinois mais en-dessous de ses concurrents sur le russe. JUSTEXT s'impose sur cette langue, ce qui semble valider la robustesse de son approche multilingue fondée sur les mots outils. Ses résultats sont compétitifs sur l'anglais et c'est sans doute sur le chinois qu'il perd la confrontation à distance avec BP3_ART. En effet les modèles langagiers de JUSTEXT sur les mots outils ne sont pas applicables à la langue chinoise. Si l'on partait des résultats sur l'anglais pour choisir un outil d'extraction de contenu, nous pourrions être tentés de choisir NEWSPAPER ou encore GOOSE. Mais leurs performances sont très variables, en particulier pour le grec (moins de 6% de F-mesure avec un rappel très faible). NEWSPAPER et NEWSPLEASE apparaissent véritablement spécialisés sur l'anglais.

Visualisation de la variation à l'échelle des documents Afin de mieux visualiser ces variations nous présentons dans la Figure 1 les résultats par document pour le corpus global pour les 6 outils les plus performants au point de vue monolingue ou multilingue. Les écarts types sont plutôt élevés en général (± 24 sur la précision pour JUSTEXT par exemple, ou sur des sous-corpus particuliers, ± 17 sur la précision de GOOSE sur l'anglais). Les graphiques permettent de saisir d'un coup d'œil l'importance de cette variabilité. On peut ainsi observer que les documents en anglais sont mieux



(a) `occ_eval` GOOSE (b) `occ_eval` JT (c) `occ_eval` READABILITY

(d) `occ_eval` NEWSPAPER (e) `occ_eval` BP3_ARTICLE (f) `occ_eval` TRAFILATURA

FIGURE 1: Visualisation mettant en rapport la précision (abscisse) et le rappel (ordonnée) pour chaque document du corpus (*el* = grec, *en* = anglais, *pl* = polonais, *ru* = russe, *zh* = chinois)

traités (en bleu). Les points correspondant au grec (en gris) sont peu nombreux, ce qui correspond à un plus faible nombre de sources (les points groupés sur la même abscisse). La dispersion des points et les codes couleur permettent ainsi de saisir des informations sur la composition du corpus lui-même, et d'en déduire ici qu'il peut être intéressant de tenir compte du nombre de documents par source ou de retenir la macro-moyenne sur les sources.

En regard de ces graphiques, le comportement des outils peut être classé en trois catégories distinctes. GOOSE est plutôt efficace en termes de précision, d'où le grand nombre de points situés à droite du graphique mais aussi les problèmes afférents bien visibles par des pics de faible rappel. Au contraire, JUSTEXT offre un bon rappel pour plus de documents, d'où le grand nombre de points en haut de la courbe. Enfin, READABILITY laisse apparaître une diagonale, ce qui suggère plus de résultats équilibrés entre la précision et le rappel et explique pour partie les bons résultats en F-mesure, et notamment en chinois (en rouge dans la courbe). TRAFILATURA se manifeste par une dispersion des points plus homogène que les autres outils, signe que l'outil n'a pas de réel point faible mais pas non plus de point fort. Cette performance est mesurée par le meilleur résultat en macro-moyenne, tandis que les problèmes de précision en anglais notamment visibles ici permettent d'expliquer la performance plus faible capturée par la micro-moyenne.

3 Quelques problèmes liés au contenu

Nous avons montré que l'extraction est en elle-même problématique et souhaitons à présent examiner à titre d'exemple des problèmes liés aux textes trouvés par des méthodes de parcours du web. Ces dernières créent un phénomène de « pêche au chalut » en ce que le taux de remplissage ainsi que la

qualité des spécimens capturés ne se voit qu'a posteriori, pouvant nécessiter un tri. La recherche et du suivi de liens de manière indiscriminée ou non-supervisée implique que des sites comportant une forte concentration d'interliens seront « (re)pêchés » plus rapidement que d'autres. La question de l'échantillonnage renvoie à des problèmes existants qui ne feront qu'indirectement l'objet de cette étude, à travers des textes et communautés problématiques. De même, les méthodes d'optimisation des résultats d'une page sur les moteurs de recherche (*search engine optimization*) avec des textes modifiés ou même générés automatiquement ne sont pas examinés mais entrent en résonance avec les problématiques que nous décrivons. Enfin, la publicité, dissimulée ou non (sur les blogs de mode par exemple), fait également figure de limitation à l'approche opportuniste. Afin de fournir des exemples concrets des biais que l'on peut observer, nous proposons une typologie des discours de haine et des potentielles infractions en termes légaux, qui peuvent s'appliquer par extension aux textes republiés par un projet de recherche. Les exemples cités ci-dessous sont problématiques et pourraient heurter la sensibilité des lecteurs et lectrices.

Notre examen qualitatif s'appuie sur un corpus de documents en allemand établi comme base empirique au sein du projet de lexicographie DWDS (Geyken *et al.*, 2017) [6]. Ces documents web proviennent majoritairement d'Allemagne, mais aussi d'Autriche et de Suisse. Dans ce contexte, les interliens font sens pour des communautés distinctes du reste du web (par exemple les bloggeu·r·se·s de mode en Autriche) mais exposent également à des risques en déformant ou en rattachant le contenu du corpus à un genre ou un groupe précis. La propagande d'extrême-droite représente d'après nos observations une portion significative des documents problématiques. Nous n'avons pas trouvé de cas avérés contraires à la loi sur le reste du spectre politique, si ce n'est un exemple unique en son genre : une page de propagande nord-coréenne rédigée en allemand. Dans la plupart des cas nous ne proposerons pas de liens vers les documents, d'une part à cause d'un problème strictement légal (apologie de crimes) mais également pour des raisons éthiques (ne pas servir de florilège). Les documents restant après filtrage peuvent être interrogés en ligne.

3.1 Description

Appareil législatif Confronté à des abus ou irrégularités, le législateur désire souvent réguler les discours de haine. Nous pouvons évoquer les cas suivants, tous formalisés en droit : (1) par référence à l'anti-constitutionnalité de certains groupes et ce qui en découle pour les documents et les textes produits, par exemple les lois contre la propagande fasciste ; (2) en raison d'une incitation caractérisée à la haine raciale, infraction pénale depuis 1972 en droit français, loi revue en 2015 en Allemagne ; (3) des cas clairement établis en droit, comme la négation de la Shoah, la négation de la culpabilité dans les crimes commis pendant la période nationale-socialiste ou le révisionnisme ; (4) la description et apologie de la violence, notamment sous l'angle de la protection de la jeunesse : déclarations et slogans contraires aux droits de l'Homme, bellicisme.

Remises en cause systémiques Outre des slogans exploités par le personnel politique d'extrême-droite (« on n'est plus chez soi »), la remise en cause d'un soi-disant « système » est un élément central des textes problématiques du corpus, qui se font parfois injurieux, par exemple dans le cas de la critique d'un soi-disant consensus tourné vers les médias mainstream (*Mainstream-Medien*), où des « putes journaleuses » (*Medienhuren*) censureraient certaines idées.

6. https ://www.dwds.de

Sexisme et racisme Sexisme et racisme sont des corollaires parfois sous-jacents mais omniprésents des théories complotistes et des remises en cause systémiques, comme à travers des discours ouvertement xénophobes sur le soi-disant afflux ou trop-plein d'étrangers ainsi que la dénonciation d'une soi-disant « fémocratie », pouvoir féministe et injustement castrateur qu'il conviendrait d'identifier, de brider, voire d'éliminer.

La pornographie constitue un cas à part, tant il s'agit d'une industrie majeure de production et de publication de contenu, qui génère un trafic considérable. Par conséquent, toute collecte de données va trouver des hôtes hébergeant des annonces ou des vidéos, comportant nettement plus de pages web que d'autres ainsi que des liens pour favoriser le référencement. L'impact sur des corpus est réel, dans une collecte ciblant des sites utilisant WordPress, décrite dans Barbaresi (2016), *mydirtyhobby* (nom de marque) figure parmi les catégories et tags les plus fréquents des sites en `.at` (Autriche). La présence de descriptions de vidéos pornographiques dans un corpus, si elle paraît logique dans une certaine proportion au vu de la large diffusion de ces sites, a un fort impact au niveau lexical, avec un vocabulaire mais aussi un imaginaire issus du règne animal (docilité attendue des « femelles », impétuosité des « mâles ») et des concepts souvent hétérosexuels et dominateurs comme le mot composé *Dreilochstute* (« jument à trois trous »), mot quasiment absent des corpus de référence et plusieurs centaines de fois plus fréquent dans les corpus web non-filtrés.

Apologie du fascisme ou du national-socialisme L'apologie de crimes passés se fait notamment à travers des mots d'ordre de la période national-socialiste, qui servent de moyen d'identification (« *Deutschland erwache* », « *Allemagne réveille-toi* »), ce qui modifie en conséquence la teneur du corpus. Les documents connexes utilisent parfois une iconographie nationale-socialiste glorifiant les principales figures du régime qui pose un problème d'identification puisqu'elle est parfois introuvable dans le texte. En effet, l'idolâtrie par le biais d'images ne semble pas aussi strictement poursuivie.

Négationnisme Révisionnisme et négationnisme portent notamment sur une remise en cause ou une discussion du nombre de personnes enfermées et exécutées pendant la période nazie. Face à la répression de ces discours qui tombent sous le coup de la loi, les groupes concernés semblent opérer par mots-clés, comme celui de température des chambres à gaz (*Gaskammertemperatur*), concept désignant une théorie (absolument fausse) visant à exonérer la hiérarchie des camps de concentration de toute responsabilité dans l'extermination de millions de personnes.

Théories conspirationnistes Un bon indicateur de théories conspirationnistes consiste à chercher le néologisme *Reptiloiden* / reptiliens ainsi que des thématiques connexes pour débusquer des documents problématiques. Si le concept de créatures imaginaires (ici des reptiles à figure humaine) ayant pris le pouvoir ou le contrôle des gouvernants peut sembler inoffensive, il s'agit bien d'opérer une distinction entre des êtres humains et des nuisibles, catégories qui en recouvrent d'autres ou opèrent de manière souple pour rappeler d'autres théories du complot, par exemple à travers le syllogisme « X (forme attestée dans le corpus : Angela Merkel) est un reptile », « X poursuit une œuvre secrète de destruction ou d'accaparement des ressources », « les reptiles sont des juifs ».

3.2 Conséquences

Filtrer les documents problématiques sans fausser l'échantillon prélevé sur le web représente un triple problème :

éthique corpus et outils de recherche se transforment en un raccourci vers des discours d'extrême-droite, ils peuvent ou doivent être utilisés pour effectuer des signalements et constituer des corpus spécialisés utiles notamment en sciences politiques ou en sociologie ;

légal certaines pages tombent clairement sous le coup de la loi et nécessitent un dépistage et une intervention immédiate, d'autres pas nécessairement tout en présentant un risque difficilement appréciable pour des néophytes ;

linguistique un compromis en forme de « ni-ni » paraît adéquat : ni conserver en l'état, ni supprimer tous les documents ou les occurrences. Malgré la distorsion des corpus, garder des échantillons portant trace de différents types de discours permet d'offrir une perspective large et à l'image de l'époque sur les modes d'expression en ligne.

4 Conclusions

La collecte et l'usage de données web sont sujets à une série de problèmes éthiques, méthodologiques et épistémologiques qui méritent l'attention de la communauté scientifique. Il appert que les approches opportunistes présidant à l'établissement de grands corpus tirés du web ne sont pas sans poser un certain nombre de difficultés. Nous avons apporté des preuves empiriques de leur impact, tout d'abord en étudiant la forme des documents obtenus à travers la comparaison de méthodes d'extraction des données et ensuite en recensant des problèmes centrés sur le contenu des corpus et liés aux méthodes d'acquisition opportuniste des données. La faible supervision conduit à un *far west*, *« Wild West Web Crawling »* selon Jo & Gebru (2020), tandis qu'une approche plus supervisée et maîtrisée ne suffit pas à résoudre des problèmes posés par l'extraction de texte.

Au phénomène de dispersion des segments textuels visible sur les graphiques d'évaluation répond une probabilité élevée de cerner certaines communautés (hobby précis ou frange politique) et genres textuels (petites annonces et annuaires). Sur la forme, les corpus web peuvent receler des documents incomplets et tronqués ainsi que des doublons et des segments génériques, dans une proportion variable qui pourrait bien être inconnue ou mal estimée par la communauté scientifique. Par ailleurs, des problèmes de fond substantiels peuvent surgir. Des textes ou éléments indésirables se trouvent dans des corpus destinés à la recherche en linguistique et en TAL, d'une part à cause de l'impossible contrôle des sources et adaptation à certains types de pages dès que la taille du corpus atteint un certain ordre de grandeur, et d'autre part en raison de l'application d'outils génériques et supposés adéquats sans vérification de leur efficacité pour des textes, langues ou sujets divergents, problématique connue en apprentissage artificiel par la notion d'adaptation de domaine.

Il faudrait pouvoir non seulement décrire ces problèmes mais également les circonscrire, ce qui implique de trouver des méthodes de mesure ainsi que des heuristiques de limitation. Alors que le détourage peut être évalué et résolu par des approches quantitatives (comprenant étalons et métriques), les difficultés d'ordre qualitatif sont plus difficile à cerner et à étalonner, alors même que leurs potentielles conséquences éthiques voire pénales sont plus graves. La « corne d'abondance » représentée par la collecte de données massives à coût moindre semble bien réelle mais est en réalité assujettie à un examen approfondi en termes de calibrage et d'équilibrage afin de constituer une nécessaire assise scientifique et de dépasser les logiques opportunistes.

Références

BARBARESI A. (2015). *Ad hoc and general-purpose corpus construction from web sources*. Thèse de doctorat, École Normale Supérieure de Lyon.

BARBARESI A. (2016). Efficient construction of metadata-enhanced web corpora. In P. COOK, S. EVERT, R. SCHÄFER & E. STEMLE, Éds., *Proceedings of the 10th Web as Corpus Workshop*, p. 7–16 : Association for Computational Linguistics.

BARBARESI A. (2019). Generic Web Content Extraction with Open-Source Software. In *Proceedings of KONVENS 2019, Kaleidoscope Abstracts*, p. 267–268 : GSCL.

BARBARESI A. & LEJEUNE G. (2020). Out-of-the-Box and Into the Ditch ? Multilingual Evaluation of Generic Text Extraction Tools. In *Proceedings of the 12th Web as Corpus workshop (WAC-XII)* : ELRA. à paraître.

BARONI M., CHANTREE F., KILGARRIFF A. & SHAROFF S. (2008). Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*, p. 638–643 : ELRA.

CALBERAC Y. (2010). *Terrains de géographes, géographes de terrain. Communauté et imaginaire disciplinaires au miroir des pratiques de terrain des géographes français du XXe siècle*. Thèse de doctorat, Université Lumière Lyon 2.

CALISKAN A., BRYSON J. J. & NARAYANAN A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

GEYKEN A., BARBARESI A., DIDAKOWSKI J., JURISH B., WIEGAND F. & LEMNITZER L. (2017). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, **45**(2), 327–344.

HAMBORG F., MEUSCHKE N., BREITINGER C. & GIPP B. (2017). news-please : A generic news crawler and extractor. In M. GAEDE, V. TRKULJA & V. PETRA, Éds., *Proceedings of the 15th International Symposium of Information Science*, p. 218–223.

JO E. S. & GEBRU T. (2020). Lessons from Archives : Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 306–316.

KOHLSCHÜTTER C., FANKHAUSER P. & NEJDL W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, p. 441–450.

LEJEUNE G. & BARBARESI A. (2020). Bien choisir son outil d'extraction de contenu à partir du Web. In *Actes de la conférence JEP-TALN-RECITAL 2020, Démonstrations* : ATALA. à paraître.

LEJEUNE G. & ZHU L. (2018). A New Proposal for Evaluating Web Page Cleaning Tools. *Computación y Sistemas*, **22**(4).

MCENERY T. & HARDIE A. (2011). *Corpus linguistics : Method, theory and practice*. Cambridge University Press.

OLSTON C. & NAJORK M. (2010). Web Crawling. *Foundations and Trends in Information Retrieval*, **4**(3), 175–246.

PETERS M. E. & LECOCQ D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22nd International Conference on World Wide Web*, p. 89–90.

POMIKÁLEK J. (2011). *Removing boilerplate and duplicate content from web corpora*. Thèse de doctorat, Masaryk University.

RATCLIFF J. W. & METZENER D. E. (1988). Pattern Matching : The Gestalt Approach. *Dr. Dobb's Journal*, **13**(7), 46.

SCHÄFER R., BARBARESI A. & BILDHAUER F. (2013). The Good, the Bad, and the Hazy : Design Decisions in Web Corpus Construction. In *Proceedings of the 8th Web as Corpus Workshop*, p. 7–15.

SMITH J., SAINT-AMAND H., PLAMADĂ M., KOEHN P., CALLISON-BURCH C. & LOPEZ A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, p. 1374–1383.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Challenges in the Management of Large Corpora (CMLC-7) 2019*, p. 9–16.

TANGUY L. (2013). La ruée linguistique vers le Web. *Texto ! Textes et Cultures*, **18**(4).

VALETTE M. (2008). Pour une science des textes instrumentée. *Syntaxe et sémantique*, **9**, 9–14.

WENINGER T., PALACIOS R., CRESCENZI V., GOTTRON T. & MERIALDO P. (2016). Web Content Extraction : A Meta-Analysis of Its Past and Thoughts on Its Future. *SIGKDD Explorations Newsletter*, **17**(2), 17–23. DOI : 10.1145/2897350.2897353.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Répliquer et étendre pour l'alsacien « Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux »

Alice Millour[1] Karën Fort[1, 2] Pierre Magistry[3]
(1) Sorbonne Université / STIH, 28, rue Serpente 75006 Paris, France,
(2) Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
(3) Aix-Marseille Université, ENP-China, Irasia, 13100 Aix-en-Provence, France
alice.millour@sorbonne-universite.fr, karen.fort@sorbonne-universite.fr,
pierre.magistry@univ-amu.fr

RÉSUMÉ

Nous présentons ici les résultats d'un travail de réplication et d'extension pour l'alsacien d'une expérience concernant l'étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux (Magistry *et al.*, 2018). Ce travail a été réalisé en étroite collaboration avec les auteurs de l'article d'origine. Cette interaction riche nous a permis de mettre au jour les éléments manquants dans la présentation de l'expérience, de les compléter, et d'étendre la recherche à la robustesse à la variation.

ABSTRACT

Replicating and extending for Alsatian : "POS tagging for low-resource languages by adapting word embeddings"

We present here the results of our efforts in replicating and extending for Alsatian an experiment concerning the POS tagging of low-resourced languages by adapting word embeddings (Magistry *et al.*, 2018). This work was performed in close collaboration with the authors of the original article. This rich interaction allowed us to identify the missing elements in the presentation of the experiment, to add them and to extend the experiment to robustness to variation.

MOTS-CLÉS : réplicabilité, étiquetage en parties du discours, langues peu dotées, variation.

KEYWORDS: replicability, POS-tagging, low-resourced languages, variation.

1 Motivations

Les avancées obtenues ces dernières années en traitement automatique des langues (TAL) grâce à l'apprentissage neuronal sont largement dépendantes de la disponibilité de très gros corpus dans les langues considérées. Or, pour de très nombreuses langues (la majorité, dites « peu dotées »), de tels corpus sont inexistants. Un article présenté à TALN 2018 (Magistry *et al.*, 2018) propose une solution partielle à ce problème pour l'étiquetage automatique en parties du discours, par spécialisation des plongements lexicaux. Les auteurs y annoncent des résultats supérieurs à l'état de l'art, notamment pour l'alsacien (0,91 d'exactitude).

Les corpus utilisés et le tagger entraîné ont été développés dans le cadre du projet RESTAURE porté

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 29–37.
HAL : hal-02750224.

par D. Bernhard (LiLPa, Strasbourg). Notons que le développement parallèle des ressources et des outils a pu entraîner un certain nombre de difficultés liées à la publication asynchrone de chacun des éléments. Ceux-ci n'ont pas été publiés avec un suivi de versions strict alors que les corpus présentent des évolutions critiques, notamment la modification du jeu d'étiquettes au fur et à mesure.

Travaillant sur le sujet, nous avons souhaité reproduire cette expérience dans le but de tester la robustesse du modèle à la variation, un phénomène très répandu dans les langues peu dotées, en particulier les langues non standardisées.

Pour ne pas nous inscrire dans une logique concurrentielle, mais plutôt dans une dynamique de passage de flambeau permettant de construire des solutions véritablement ré-utilisables, nous avons choisi de bâtir notre recherche en collaboration avec les auteurs de l'article original. Nous détaillons ici le processus de réplication et les résultats que nous avons obtenus en étendant l'expérience à la robustesse à la variation.

2 Reproduire ou répliquer ?

La terminologie utilisée mérite qu'on s'y attarde, tant elle rend compte de la complexité de l'acte, apparemment simple, de rejouer une expérience décrite dans un article de recherche. Nous reprenons ici de manière succincte les questions mises au jour et détaillées dans (Cohen *et al.*, 2018). La réplicabilité est une propriété d'une expérience, celle d'être rejouée ou répétée [1], alors que la reproductibilité est une propriété des **résultats** de l'expérience menée : on peut obtenir les mêmes conclusions ou les mêmes valeurs [2]. Nous nous intéressons ici en priorité à répliquer l'expérience (pour mieux la comprendre), pour ensuite tenter d'en reproduire le résultat (pour être sûr de partir sur les mêmes bases) pour, enfin, étendre l'expérience.

De tels efforts sont de plus en plus valorisés dans le domaine du TAL. Après deux ateliers à LREC 2016 et 2018 (Branco *et al.*, 2016, 2018), une *shared task*, REPROLANG [3], a été organisée dans le cadre de LREC 2020. La possibilité de rejouer une expérience est même devenue un critère de sélection pour COLING 2018. Une étude menée parmi les chercheurs du domaine a montré que le sujet est perçu comme un problème important par la majorité des répondants (Mieskes *et al.*, 2019) et que, lorsque ceux-ci ont essayé de reproduire une expérience (et y sont parvenus), les résultats obtenus se sont très souvent révélés significativement différents de ceux publiés. Cela ne signifie pas pour autant que les auteurs originaux sont de mauvaise foi. Simplement, le manque de documentation des expérimentations empêche souvent de se replacer dans les conditions expérimentales de l'expérience initiale [4].

Parmi les éléments trop souvent mal documentés, les pré-traitements (dont la tokénisation) et les versions des logiciels et des ressources langagières utilisées sont des classiques (Fokkens *et al.*, 2013). L'expérience reproduite ici ne fait pas exception, malgré les efforts de ses auteurs.

1. "Replicability or repeatability is a property of an experiment : the ability to repeat –or not– the experiment described in a study." (p. 3)
2. "Reproducibility is a property of the outcomes of an experiment : arriving –or not– at the same conclusions, findings, or values." (p. 3).
3. Voir : https://lrec2020.lrec-conf.org/en/reprolang2020/.
4. Nous ne formulons pas ici d'hypothèse quant aux raisons méthodologiques ou pratiques de ce manque de documentation dans le domaine du TAL. Il nous a été signalé par un des relecteurs que dans d'autres domaines souffrant tout autant de la précarité de leurs chercheurs, la reproductibilité systématique des expériences est assurée, notamment grâce à l'utilisation de "carnets de recherche" pour leur documentation.

3 Faire tourner le code

Plutôt que de réimplémenter la solution proposée, nous avons essayé de retrouver les conditions initiales dans lesquelles l'expérience avait été menée, tant au niveau du logiciel que des ressources langagières. Dans cette section, nous présentons donc la méthodologie que nous avons souhaitée reproduire ainsi que les éléments relatifs (i) à la disponibilité du code source et (ii) à la mise en place des configurations logicielles nécessaires pour faire tourner ce code.

3.1 Méthodologie

La méthodologie proposée permet de spécialiser les plongements lexicaux à la tâche d'annotation en parties du discours en combinant l'analyse au niveau caractère et l'utilisation des propriétés morphosyntaxiques pour un mot cible et son contexte. Le système MIMICK (Pinter *et al.*, 2017), qui se base sur la graphie des mots pour calculer les vecteurs, est utilisé pour établir les plongements des mots hors vocabulaire, nombreux dans le cas de langues peu dotées et non standardisées. Cette méthodologie peut être découpée en trois étapes permettant d'obtenir des résultats intermédiaires : i) entraînement sur un corpus brut permettant de produire un fichier de plongements lexicaux dits morphosyntaxiques, ii) entraînement du modèle de *tagger* basé sur un Bi-LSTM en utilisant les plongements lexicaux et iii) évaluation du *tagger*. Il est à noter que les deux éléments intermédiaires (fichiers de plongements et modèle du Bi-LSTM entraînés pour l'alsacien) ne sont pas distribués.

3.2 Accès au code source

Le code tel qu'utilisé dans l'expérience originale n'a pas pu être retrouvé. Le co-auteur en charge des expériences ayant terminé le postdoctorat qu'il réalisait à l'époque de la publication de l'article, il n'a aujourd'hui plus accès aux machines sur lesquelles celui-ci était stocké. Nous avons néanmoins eu accès à deux versions ultérieures du code source, correspondant à deux implémentations de la méthodologie décrite dans l'article.

Le premier code source auquel nous avons eu accès, CS_1 [5] a été mis à disposition par le premier auteur de l'article. Le dépot GitHub transmis contient une réécriture partielle du code original. Cette réécriture ayant été abandonnée avant son terme, l'ensemble des étapes réalisées dans l'expérience initiale n'y sont pas représentées.

Un second dépôt, CS_2 [6], contenant le code complet en python a été identifié dans un second temps. Il s'agit de la version simplifiée et documentée du code original, réalisée et distribuée par une postdoctorante ne faisant pas partie des auteurs initiaux de l'article.

Ces deux dépôts GitHub n'étant pas renseignés dans l'article, ni associés aux noms des auteurs, ils ne pouvaient pas être identifiés sans prise de contact avec ceux-ci. Or, ces derniers sont encore précaires et leurs affiliations changent régulièrement, nous avons donc eu de la chance d'une part de parvenir à rentrer en contact avec l'un d'entre eux malgré la désactivation de sa boîte mail, et d'autre part de pouvoir accéder à la deuxième version du code.

5. Accessible ici : https://github.com/a-tsioh/MSETagger.
6. Accessible ici : https://github.com/eknyazeva/MSETagger_py.

3.3 Accès à des modèles pré-entraînés

Bien que ce soit une pratique devenue courante en TAL, aucun des codes sources diffusés ne s'accompagne de modèles pré-entraînés. Ce système comporte trois étapes produisant des modèles : les plongements lexicaux initiaux, le modèle MIMICK qui permet de les compléter et les poids du Bi-LSTM de l'étiqueteur final. Les auteurs initiaux ont fait le choix de diffuser le code permettant de reconstruire ces modèles, mais aucun des résultats intermédiaires. Chacune de ces trois étapes recourt à de l'apprentissage profond qui suppose une initialisation aléatoire de grandes matrices de poids. La stabilité de ces modèles n'est pas garantie. elle a même d'autant plus de chances d'être problématique lorsque les corpus d'entraînement sont relativement petits, comme c'est le cas ici. [7]

3.4 Configuration logicielle

Les deux dépôts GitHub sont accompagnés de `README` contenant la majorité des informations de configuration nécessaires à l'exécution du code, notamment une liste de dépendances quasi complète. Les versions de certaines librairies python sont absentes de la documentation, mais les versions compatibles entre elles des différentes librairies ont pu être déduites à tâtons.

De la même manière, l'architecture de Bi-LSTM sur laquelle s'appuie le travail des auteurs est l'implémentation YASET (Tourille *et al.*, 2017). La version de YASET utilisée n'était précisée que dans l'un des dépôts. Dans les deux cas, lorsque les hyper-paramètres n'étaient pas précisés dans l'article, ils étaient donnés dans un fichier de configuration distribué avec le code source.

Ces difficultés, associées à la méconnaissance initiale des technologies employées par les auteurs (par exemple le langage `scala`, et le moteur de production `sbt`), constituent des freins importants à la réplication de l'expérience. La mise en place de la configuration logicielle s'est donc faite en étroite collaboration avec le premier auteur de l'article d'origine.

4 Données utilisées

4.1 Généralités sur l'alsacien

L'alsacien est un terme englobant qui regroupe les langues germaniques, principalement alémaniques, parlées en Alsace et une partie de la Moselle. L'alsacien fait partie, avec les parlers alémaniques d'Allemagne et de Suisse, des langues regroupées sous le code ISO-639-3 `gsw`. Il présente des variantes à la fois dialectales et orthographiques en raison de l'absence de standard consensuel. Nous nous intéressons à la gestion en TAL de ces variantes.

Concernant les ressources brutes disponibles, la Wikipédia alémanique (code wikipédia `als`) contient des pages écrites dans 35 langues alémaniques identifiées [8]. Les pages en alsacien sont celles catégorisées `Artikel uf Elsassisch` (1 893 pages) et `Artikel uf Elsässisch` (1 page). La majorité de ces pages concernent des lieux et sont très semblables entre elles.

7. Un article plus long, décrivant le système plus en détails et détaillant ce problème était en cours de rédaction suite à l'article de TALN 2018, mais il n'a pas pu être terminé avant la fin du projet ANR.

8. Voir les catégories commençant par "`Artikel uf`", https://als.wikipedia.org/wiki/Spezial:Kategorie consultées en février 2020

4.2 Corpus bruts

Les corpus bruts utilisés pour entraîner les plongements lexicaux sont les corpus C_{Brut}_56k et C_{Brut}_200k. C_{Brut}_56k, communiqué sur demande par l'un des auteurs l'ayant lui même obtenu de D. Bernhard, est constitué d'un ensemble de 103 pages Wikipédia totalisant 56 965 tokens. Ce corpus, libre de droit, peut être reconstruit à partir de la liste des pages fournies avec le corpus. C_{Brut}_200k a été obtenu ultérieurement auprès de D. Bernhard. C'est un ensemble de documents contenant des pages de la Wikipédia alémanique rédigées en alsacien, ainsi que des documents dont les licences ne sont pas claires. La proportion de ce corpus qui est effectivement libre de droit n'a pas été déterminée.

4.3 Corpus annoté

Le corpus annoté de l'alsacien utilisé pour entraîner et évaluer le *tagger*, $C_{Annoté}$, est distribué sous licence CC BY-SA [9]. C'est un ensemble constitué de (i) pages de la Wikipédia alémanique écrites en alsacien (ii) chroniques publiées par le conseil général du département du Haut-Rhin, (iii) une recette et (iv) un extrait de pièce de théâtre, totalisant 12 644 tokens annotés (Bernhard *et al.*, 2018).

5 Résultats obtenus

Les résultats que nous obtenons diffèrent des résultats publiés précédemment. Une partie de la variation observée peut s'expliquer par la difficulté à reconstituer des corpus similaires (notamment la division en jeu d'entraînement et jeu test). Il semble aussi qu'une grande part de cette variation est à attribuer à l'instabilité des plongements lexicaux entraînés sur de petits corpus (voir Section 3.3). Ceci pose la question de l'importance de la diffusion de modèles pré-entraînés. Une telle pratique favorise la reproductibilité des résultats mais dans le même temps, elle masque des propriétés importantes de la chaîne de traitement complète.

5.1 Premières expériences, réalisées avec la réécriture partielle du code (CS_1)

La première tentative de reproduction des résultats a été réalisée à partir du code CS_1 en utilisant C_{Brut}_56k pour entraîner les plongements, 80 % de $C_{Annoté}$ pour entraîner le modèle, et 20 % $C_{Annoté}$ pour l'évaluer.

Cette expérience nous a permis d'attester que nos conditions logicielles étaient les mêmes que celles de l'auteur initial (à ce jour) : nous avons en effet mené cette expérience en parallèle et obtenu le même résultat (une exactitude du *tagger* de 0,78). Il n'y a donc pas d'élément de configuration implicite n'ayant pas été communiqué par l'auteur. En revanche, la taille du corpus C_{Brut}_56k transmis par l'auteur ne correspondant pas aux données présentées dans l'article initial, nous avons poussé nos recherches pour finalement obtenir l'accès au corpus C_{Brut}_200k.

Cette expérience a également permis de mettre au jour que soit le corpus $C_{Annoté}$ disponible à ce jour en ligne n'est pas dans l'état dans lequel les expériences initiales ont été menées, soit la réécriture du code utilisé à l'époque est incomplète et ne gère plus certains cas particuliers propres à l'alsacien et

9. Voir https://zenodo.org/record/2536041.

pris en charge avant la réécriture. Nous n'avons en effet pas pu retrouver plusieurs éléments utilisés à l'époque, tels qu'un filtre sur les corpus bruts permettant d'éliminer les entrées de dictionnaire, et une opération visant à uniformiser les jeux d'étiquettes.

Concernant le jeu d'étiquettes et la tokénisation, l'article initial ne mentionne pas les choix qui ont été faits à ce sujet. Les corpus C_{Brut} et $C_{Annoté}$ sont aujourd'hui disponibles tokénisés de deux manières différentes, et le tokéniseur distribué pour l'alsacien [10] ne gère pas les cas divergents, en l'occurrence le découpage – ou non – des contractions de prépositions (ADP) et déterminants (DET), par exemple : « *zum*/ADP+DET », découpé en « *zu*/ADP *dem*/DET ».

Nous avons réalisé une seconde expérience en utilisant C_{Brut}_200k pour entraîner les plongements, et en utilisant les mêmes corpus que précédemment après uniformisation du jeu d'étiquettes. Cette nouvelle configuration nous a permis d'obtenir un *tagger* d'une exactitude de 0,81. Un score de 0,87 a été obtenu plus tard par l'auteur de l'article après activation d'une option non spécifiée dans la documentation.

Ces diverses expériences ont donc montré que la réplicabilité du travail en question ne pouvait se faire sans que l'auteur ne complète le code mis à disposition. Par ailleurs, certaines ressources langagières, non librement disponibles, n'ont pu être retrouvées que par relations inter-personnelles. Enfin, certains traitements (en particulier la tokénisation) n'étaient pas suffisamment documentés et n'ont pas pu être reconstitués, ce qui, comme nous l'avons précisé en section 2, est un oubli classique.

5.2 Un pas plus loin : tester la robustesse à la variation avec le code CS_2

Nous avons fixé le paramètre `patience` à 75 pour toutes les expériences réalisées avec le code C_2, la réécriture en python simplifiée et documentée du code original. La première expérience réalisée, en utilisant C_{Brut}_200k pour l'entraînement des plongements, et les corpus aux jeux d'étiquettes normalisés décrits ci-dessus, nous a permis d'obtenir une exactitude de 0,89 (valeur moyenne sur un 5-fold avec un écart-type de 0,005). Ce code implémente selon nous de manière fiable la méthodologie présentée par les auteurs de l'article d'origine. Nous l'avons donc utilisé pour mener des expériences additionnelles, afin d'en tester la robustesse à la variation.

Pour ce faire, nous avons séparé le corpus annoté en deux sous-ensembles, en nous basant sur une caractéristique linguistique identifiée : la prédominance de la terminaison des noms et adjectifs en "-e" dans les variantes du nord, et en "-a" dans les variantes du sud (Brunner, 2001). Chaque sous-ensemble n'est pas uniforme et contient lui même plusieurs variantes, par exemple la variante strasbourgeoise parmi les variantes du nord.

Nous avons fait l'hypothèse qu'un article Wikipédia ne contenait qu'une seule variante. Néanmoins, lorsque le calcul des fréquences relatives des terminaisons propres à chaque variantes ne nous permettait pas de décider, nous avons examiné le fichier à la main. Dans tous les cas, nous avons pu identifier le biais à l'origine de l'équilibre des terminaisons, comme la fréquence élevée d'un élément de vocabulaire (par exemple le déterminant "***de***"), ou la présence de mots en français (par exemple, "Stad**e** d**e** l'Ill"). Nous avons ainsi pu attribuer à chaque article la variante lui correspondant. Les fréquences relatives des terminaisons en "-e" et en "-a" sont en moyenne d'un facteur 30. Nous avons ainsi pu déterminer que 40 % du corpus annoté contenait des variantes du sud ($C_{Annoté-Nord}$, 4 998 tokens) et 60 % des variantes du nord ($C_{Annoté-Sud}$, 7 646 tokens).

10. Distribué par l'équipe du projet RESTAURE, voir https://zenodo.org/record/2454993.

Les résultats présentés dans le tableau 1 ont été obtenus en découpant les deux corpus obtenus en 3 sous-corpus : corpus d'entraînement ($C80_{Annoté-X}$, 80 %), de développement (10 %), et d'évaluation ($C10_{Annoté-X}$).

	$C10_{Annoté-Nord}$	$C10_{Annoté-Sud}$
$C80_{Annoté-Nord}$	0,75	0,74
$C80_{Annoté-Sud}$	0,74	0,79

TABLE 1 – Résultats de l'entraînement sur des corpus plus uniformes quant aux variantes présentes dans les corpus d'entraînement et d'évaluation

En première analyse, et bien que les corpus d'évaluation soient de taille réduite, il semble que la méthodologie proposée soit sensible aux variantes présentes dans les corpus : les performances les meilleures sont obtenues lorsque le corpus d'entraînement et d'évaluation contiennent les mêmes variantes. Notamment, les performances du *tagger* entraîné sur le corpus $C80_{Annoté-Sud}$ diminuent de 4 points sur le corpus d'évaluation $C10_{Annoté-Nord}$. Il serait intéressant de prolonger cette étude en mesurant à taille de corpus égales pour les deux sous-variantes, l'impact de la présence - ou non - de celles-ci dans le corpus utilisé pour entraîner les plongements.

6 Conclusions et perspectives

Nous avons présenté ici une expérience de réplication collaborative d'un article publié à TALN 2018. Le logiciel conçu à l'époque n'est plus disponible en tant que tel et nous n'avons pas réussi à le reconstruire à partir des pièces accessibles au premier auteur de l'article. Nous avons appris entre temps qu'il a été repris par une autre personne (précaire également), a fait l'objet d'améliorations et est désormais disponible sur un autre dépôt GitHub [11]. Nous avons répliqué l'expérience sur cette nouvelle base (qui n'est pas non plus celle de l'article initial), mais ne sommes pas parvenus à en reproduire les résultats (0,87 avec CS_1, 0,89 avec CS_2, *vs* 0,91 dans l'article initial).

Comme évoqué en section 3.3, la question de la distribution des résultats intermédiaires (fichiers de vecteurs, modèles de *tagger*) se pose dans le cas général. Cependant, dans le contexte de langues peu dotées, la distribution de modèles instables ne paraît pas indiquée. Dans ce contexte, l'effort doit donc être dirigé en priorité vers l'accessibilité aux ressources. En effet, l'accès à ces dernières pose de nombreux problèmes : le corpus C_{Brut} à partir duquel sont entraînés les plongements dans l'expérience initiale n'est en effet pas librement disponible.

Cela pose la question de la définition d'un résultat « état de l'art ». Selon nous celle-ci devrait être précisée pour prendre en compte les cas où la ressource d'origine ne peut pas être ré-utilisée librement.

Il paraît évident que la reproductibilité d'une expérience ne peut être assurée sans une documentation précise des conditions de l'expérience au sens large mais aussi du protocole d'évaluation. Il nous semble que la pratiques à mettre en œuvre dépendent largement de l'expérience et de la méthode d'évaluation initiale, c'est pourquoi nous nous gardons de dresser ici une liste de bonnes pratiques à partir d'une seule tentative de reproduction d'expérience.

Enfin, une partie des obstacles que nous avons rencontrés est liée au cadre dans lequel la recherche est réalisée. Parmi les difficultés rencontrées figurent en effet l'instabilité des affiliations des chercheurs,

11. Voir : https://github.com/eknyazeva/MSETagger_py.

l'urgence dans laquelle les recherches sont réalisées, l'abandon des démarches de pérennisation du code, etc. Comme il a été proposé par plusieurs relecteurs, ils serait intéressant de mener une étude comparative des pratiques de documentation dans différentes disciplines souffrant tout autant de la précarité de leurs chercheurs.

Remerciements

Nous remercions A-L. Ligozat et S. Rosset (LIMSI-CNRS) ainsi que D. Bernhard (LiLPa, Strasbourg) pour leur disponibilité, leurs conseils et l'aide qu'elles nous ont apporté. Nous remercions également les relecteurs de l'atelier ETeRNAL, qui nous ont permis, grâce à leurs remarques constructives, d'améliorer notre article.

Références

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLE L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In *Actes de 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japon. HAL : hal-01704806.

BRANCO A., CALZOLARI N. & CHOUKRI K., Éds. (2016). *Proceedings of the Workshop on Research Results Reproducibility and Resources Citation in Science and Technology Proceedings*.

BRANCO A., CALZOLARY N. & CHOUKRI K., Éds. (2018). *Proceedings of the 4REAL 2018 - Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, Paris, France. European Language Resources Association.

BRUNNER J.-J. (2001). *L'alsacien sans peine*. Assimil.

COHEN K. B., XIA J., ZWEIGENBAUM P., CALLAHAN T., HARGRAVES O., GOSS F., IDE N., NÉVÉOL A., GROUIN C. & HUNTER L. E. (2018). Three Dimensions of Reproducibility in Natural Language Processing. In N. C. C. CHAIR), K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Actes de the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

FOKKENS A., VAN ERP M., POSTMA M., PEDERSEN T., VOSSEN P. & FREIRE N. (2013). Offspring from Reproduction Problems : What Replication Failure Teaches Us. In *Actes de the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, p. 1691–1701.

MAGISTRY P., LIGOZAT A.-L. & ROSSET S. (2018). Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. In *Actes de Conférence sur le Traitement Automatique des Langues Naturelles (TALN'18)*, Rennes, France. HAL : hal-01793092.

MIESKES M., FORT K., NÉVÉOL A., GROUIN C. & COHEN K. B. (2019). NLP Community Perspectives on Replicability. In *Recent Advances in Natural Language Processing*, Varna, Bulgarie. HAL : hal-02282794.

PINTER Y., GUTHRIE R. & EISENSTEIN J. (2017). Mimicking word embeddings using subword RNNs. In *Actes de the 2017 Conference on Empirical Methods in Natural Language Proces-*

sing, p. 102–112, Copenhague, Danemark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1010.

TOURILLE J., FERRET O., NÉVÉOL A. & TANNIER X. (2017). Neural architecture for temporal relation extraction : A bi-LSTM approach for detecting narrative containers. In *Actes de the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 224–230, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-2035.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

1990-2020 : retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire

Jean-François Bonastre[1,2]
(1) LIA, Avignon Université, 339 chemin des meinajariès, 84911 Avignon, France
(2) Association Francophone de la Communication Parlée
jean-francois.bonastre@univ-avignon.fr

RÉSUMÉ

Des enregistrements de voix se trouvent de plus en plus souvent au cœur d'affaires judiciaires importantes, notamment de par l'essor de la téléphonie mobile. La justice demande à ce que des expertises en identification de voix soient réalisées alors que dans le même temps, la pertinence scientifique de telles expertises est fortement mise en cause par les scientifiques. Ainsi, dès 1990, les chercheurs en communication parlée réunis dans le GFCP, devenu depuis AFCP, ont voté une motion affirmant que « l'identification d'un individu par sa voix est à l'heure actuelle un problème à sa connaissance non résolu ». Cette motion est toujours en vigueur, après avoir été réaffirmée en 1997 et renforcée par une pétition en 2002. Malgré cela, des expertises judiciaires en identification de voix sont réalisées en France chaque année. Cet article revient sur les actions menées par le GFCP et l'AFCP depuis la motion initiale jusqu'aux actions contemporaines. Il se propose d'évaluer les répercussions de ces actions, tant au niveau de la Justice qu'au niveau académique.

ABSTRACT

1990-2020: A look back at 30 years of discussions on voice identification in the judicial system.

Voice recordings are more and more often at the heart of important legal files, in particular due to the boom in mobile telephony. Justice demands that forensic voice identification be performed while at the same time, the scientific relevance of such expertises is widely questioned by scientists. In 1990, researchers in speech communication represented by the GFCP, which has since become the AFCP, voted a motion affirming that "the identification of an individual by his voice is currently an unsolved problem to our knowledge". This motion is still in force, after being reaffirmed in 1997 and reinforced by a petition in 2002. But forensic voice identification is still carried out each year in France. This article reviews the actions taken by the GFCP and the AFCP from the initial motion to the present day. It intends to assess the repercussions of this actions, both in terms of Justice and at the academic level.

MOTS-CLÉS : identification vocale, expertise judiciaire, identification du locuteur, fiabilité.
KEYWORDS: forensic voice identification, speaker identification, reliability.

1 Introduction

Le nombre des procès judiciaires dans lesquels des prélèvements vocaux sont présentés comme élément de preuve a connu une croissance forte durant les dernières décennies, en relation avec le

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 38–47.
HAL : hal-02750225.

taux de pénétration de la téléphonie mobile (+95 % des Français possèdent un téléphone mobile en 2020[1]). Ces prélèvements sont majoritairement issus d'interceptions téléphoniques réalisées par les enquêteurs. Lorsqu'un suspect ne reconnaît pas avoir prononcé un échantillon de voix, cas fréquent, la Justice peut demander à ce que des expertises judiciaires en identification de voix soient réalisées pour pouvoir exploiter pleinement les indices recueillis.

En réponse à ce besoin, des expertises en identification de voix sont réalisées chaque année en France depuis plusieurs décennies. Cependant, depuis tout aussi longtemps, les chercheurs académiques dans ce domaine contestent les fondements scientifiques de telles expertises. Cet article retrace leurs actions et questionne celles-ci à travers le double prisme du rôle des scientifiques et de la science dans les tribunaux.

En 1990, les scientifiques spécialistes de la parole, réunis dans le Groupe de la Communication Parlée (GCP, groupe de la Société Française d'Acoustique), ont pris officiellement position dans ce débat en votant une motion[2] affirmant notamment que « l'identification d'un individu par sa voix est à l'heure actuelle un problème à sa connaissance non résolu »[3]. Cette motion demandait également à ce que « par souci déontologique [...] tout spécialiste démontre sa compétence […] avant d'accepter de procéder à une quelconque expertise (policière, judiciaire...) ». Cette motion a été réaffirmée en 1997, puis renforcée par une pétition demandant un moratoire sur l'utilisation des expertises vocales par la Justice tant que celles-ci « n'auront été validées scientifiquement ». Les scientifiques francophones de la communication parlée, sous l'égide du GCP, devenu entre-temps GFCP puis, en 2002, l'AFCP[4] ont largement fait écho de cette prise de position, au niveau national comme au niveau international. Cet engagement s'est poursuivi dans les tribunaux en France, en revendiquant une position de *témoin scientifique* et non d'expert judiciaire. Ces actions inscrites sur la durée ont eu un fort impact sur la communauté scientifique, le milieu de l'expertise en identification de voix et le monde judiciaire, avec notamment une influence probable sur le déroulé d'un nombre significatif d'affaires judiciaires. Cet article se propose de dessiner en quelques points forts ces 30 années d'histoire, montrant ainsi le chemin parcouru, sans occulter les questions déontologiques et éthiques.

2 Le point de départ : la motion de 1990

La motion de 1990 a été motivée par l'annonce parue au bulletin officiel du 16 Octobre 1989 d'un appel à projet lancé par le ministère de l'Intérieur pour « Étude, mise au point et présentation de moyens permettant une identification de locuteur par des méthodes de comparaison à partir d'enregistrements magnétiques », le titulaire devant « mettre au point un système présentant des taux de reconnaissance aussi élevés que possible » (Boë 1999, 2000). Les laboratoires de l'AFCP ont rapidement pris conscience des dangers potentiels de cet appel à projet. Pour éviter que joue l'effet d'aubaine, ils ont décidé de communiquer au Ministère qu'aucun d'entre eux ne répondrait et de rédiger une motion unanime expliquant ce choix. Celle-ci met en avant trois faiblesses, demande un moratoire et offre les services des scientifiques pour progresser :

1 Source statista.com
2 www.afcp-parole.org/doc/MOTION_1990.pdf
3 Dans ce document, les « » indiquent des extraits littéraux.
4 Par soucis de simplicité, dans la suite de cet article, AFCP sera utilisé pour représenter (suivant l'année en question) GCP, GFCP ou AFCP.

- Insuffisance des connaissances scientifiques pour une utilisation pratique de l'identification par la voix : « L'état actuel des connaissances [ne permet pas d'identifier] un locuteur, par des procédures mises en œuvre par un expert par une méthode automatique [et cela même dans des conditions optimales]. Les travaux publiés jusqu'ici indiquent que l'expertise auditive directe n'est pas fiable, pas plus que l'examen visuel ou automatique de résultats d'analyse de la voix (spectrogrammes, etc.) ».
- Difficulté des cas réels : Les « scores obtenus en laboratoire sont [...] aggravés en situation de terrain, quand il s'agit d'identifier un locuteur sur un bref enregistrement [...] accompagné de bruit de fond, à supposer même qu'il ne cherche pas à déformer sa voix ».
- Évaluation des limites : « Il semble fondamental [...] de bien définir et de préciser – dans les conditions optimales – quelles sont les limites actuelles des méthodes scientifiques [...] des systèmes d'identification du locuteur ».
- Moratoire : « En particulier, par souci déontologique, il conviendrait que tout spécialiste démontre sa compétence en identification du locuteur avant d'accepter de procéder à une quelconque expertise (policière, judiciaire…) ».
- Ouverture : Les membres de l'AFCP « sont évidemment prêts à apporter leurs compétences dans tout projet de recherche qui puisse faire avancer les connaissances dans ce domaine, […] à contribuer à toute recherche, constitution de bases de données, expérimentation [...] en proposant des procédures rigoureuses d'évaluation [et à] participer à l'élaboration de protocoles, à des évaluations d'experts, de logiciels ou de matériels qui se présenteraient pour résoudre les problèmes d'identification juridique ».

L'espoir des chercheurs de l'AFCP était qu'à la place d'une solution opérationnelle à court terme, dangereuse et éthiquement douteuse, le Ministère finance des travaux fondamentaux, tant sur l'identification par la voix que sur l'évaluation des performances. Mais le LMS[5], un laboratoire éloigné thématiquement de l'AFCP (car travaillant sur les matériaux et la mesure), a cependant candidaté et a été retenu par le Ministère. Ce laboratoire sans compétences reconnues sur le sujet a ainsi mené des travaux sur la question de l'identification par la voix à l'aide d'une subvention publique, en contradiction avec l'avis exprimé en commun par l'ensemble des laboratoires du secteur scientifique concerné, et sans tenter d'échanger avec eux. La société Micro Surface, issue du LMS, proposera rapidement REVAO (Reconnaissance Vocale Assistée par Ordinateur), une solution (brevetée) pour l'expertise en identification de voix. Cette solution sera tout aussi rapidement controversée et rejetée par la Justice elle-même[6] (en 1992, dans le cadre de l'affaire Grégory, Boë 2000).

3 De la motion académique aux tribunaux

Le ministère de l'Intérieur a cependant pris note de la motion et, en 1992, a sollicité l'organisation d'une rencontre tripartite réunissant des représentants de la police, de la magistrature et de la recherche publique. Cette rencontre a montré une convergence[7] entre les trois communautés

[5] Le Laboratoire de Microanalyse des Surfaces relevant de l'École Nationale Supérieure de Mécanique et des Microtechniques (ENSMM)

[6] Cela n'empêchera pas la gérante de Micro Surface, Dalloul Wehbi, d'être embauchée par un laboratoire national de police scientifique et de réaliser des dizaines d'expertises judiciaires.

[7] À l'exception de D. Wehbi, qui a clamé que sa méthode REVAO, était parfaitement fiable.

concernant les doutes sur la fiabilité de l'expertise vocale et a permis de poser les premiers jalons d'une collaboration entre chercheurs et membres de la police scientifique[8].

La motion a également touché les avocats ce qui a entrainé un changement majeur de l'action de l'AFCP : les avocats ont sollicité directement les chercheurs académiques pour intervenir dans différents procès en cours, dans lesquels des expertises en identification de voix étaient produites. Le choix de l'AFCP, en conformité avec la motion, a été de refuser d'intervenir comme expert judiciaire mais d'accepter d'intervenir comme *témoin scientifique* ou *sachant*, ce qui est toujours la modalité en vigueur en 2020. Ces interventions sont en majorité réalisées à la demande de la défense et arrivent souvent au moment d'un procès en appel, généralement dans des affaires complexes, avec une forte tension. Durant les premières années, ces interventions ont été souvent perçues par le monde judiciaire comme un artifice des avocats de la défense pour amener du doute, une perte de temps dans un lieu où celui-ci est extrêmement compté. Il faut aussi reconnaître que les premiers témoignages détonnaient certainement dans l'ambiance du tribunal : comprendre le mode de fonctionnement d'un tribunal demande un vrai apprentissage et du temps. Témoigner à la barre (sans support) n'est pas donner une conférence. Gérer la violence des débats, les mises en causes personnelles et les tentatives de déstabilisation s'apprend. Mais la difficulté première reste d'expliquer, de transmettre des éléments complexes à des non spécialistes, de chercher à les convaincre sans pour autant utiliser d'artifices oratoires ou une simple analogie inadaptée. La charge émotionnelle et le poids de la responsabilité sont omniprésents tout au long de cet échange de questions et réponses souvent vif. Intervenir dans le strict cadre des fonctions académiques, en refusant tout honoraire et en se limitant aux aspects scientifiques de l'expertise[9], a été et reste un point clé pour faire reconnaître cette position de *sachant* et favoriser l'acceptabilité des témoignages. Il fait peu de doute que l'expérience du tribunal aura marqué profondément chacun des *témoins scientifiques* mandatés par l'AFCP[10].

4 Du message au moratoire

Constatant que les actions précédentes n'avaient pas fait significativement baisser le nombre des expertises judiciaires en France, l'AFCP a réaffirmé sa motion à travers une pétition[11] « pour l'arrêt des expertises vocales tant qu'elles n'auront pas été validées scientifiquement » en 1997, suite à l'affaire Prieto[12]. Celle-ci avait montré le manque de maturité des méthodes alors employées et, parfois, les insuffisances de formation en parole des experts. Bien que n'apportant aucune information complémentaire, la pétition a été mieux reçue dans le milieu judiciaire et la presse et s'est trouvée citée plus souvent durant les auditions au tribunal que la motion elle-même. Outre le fort soutien recueilli à cette occasion, son titre est certainement une clé de son succès. Il livre en quelques mots l'essentiel du message, la nécessité d'un moratoire complet sur l'expertise en identification de voix. La pétition permettra de ne plus avoir de chercheurs académiques se

---

8 Par la constitution d'un groupe de travail sur la caractérisation du locuteur et de la langue (GT1 du GDR/PRC-CHM)

9 Soit en évitant de parler l'expert en personne et de prendre position sur l'affaire en cours.

10 La majorité des témoignages a été effectuée par L.-J. Boë et J.-F. Bonastre. Les autres intervenants sont F. Bimbot, P. Dupont, P. Perrier et C. Meunier, sans oublier la participation de C. Legros, en tant que Président de la SFA.

11 http://www.afcp-parole.org/doc/petition.pdf

12 L'affaire Prieto (Boë 2000) est une affaire de terrorisme dans laquelle l'identification par la voix a joué un rôle crucial. Voir https://www.liberation.fr/societe/1999/06/16/que-valent-les-paroles-des-experts-face-a-la-voix-de-prieto-les-expertises-acoustiques-sur-la-sellet_277612

présentant comme experts judiciaires en identification de voix. Il s'agit d'un résultat fort qui a aidé à expliquer dans les tribunaux ce qu'est une *expertise scientifique*[13], offrant une objectivité assise sur une méthodologie scientifique et des travaux publiés et reconnus.

5 La nature n'aime pas le vide : du moratoire au charlatanisme

Mais la nature n'aime pas le vide… Le corollaire imprévu du moratoire a été d'inciter un individu, voyant le terrain libre, à s'attaquer au marché de l'expertise judiciaire en identification par la voix. Cette personne s'est facilement faite inscrire comme expert judiciaire (sans avoir de compétences spécifiques, ce qui illustre une faille connue du système légal français, concernant l'inscription des experts) et a créé une entreprise dédiée à cette activité, le LIPSADON (qui bénéficiera de financements publics). La Justice, face à son besoin réel, a rapidement entendu le discours rassurant et surtout largement médiatisé de ce nouvel acteur. Sans autre évaluation de ses prétentions, elle a commencé à travailler avec lui. Par le bouche à oreille, l'appel à ce prestataire s'est largement généralisé dans toute la France. Avec quelques années de décalage[14], les chercheurs en communication parlée ont été confrontés à cet individu et à une nouvelle difficulté : celle de contrer des intervenants sans aucune base scientifique ni déontologie, capables de tout, tant dans les rapports d'expertise que durant les témoignages. Les affrontements ont été particulièrement difficiles, avec des confrontations directes de personne à personne. Ils ont cependant permis de montrer les limites de ce prestataire et, petit à petit, les magistrats ont retransmis les doutes profonds sur les méthodes employées et la déontologie du LIPSADON. Mais les représentants de l'AFCP ont dû pour cela mettre en cause l'expert lui-même (voir Boë 2012a et Boë 2012b). Cela nécessite un engagement différent, avec en retour des attaques personnelles et des questionnements personnels également (mettre en cause publiquement un individu, sa probité, n'équivaut pas à évaluer les aspects scientifiques d'une expertise). Le LIPSADON et son directeur (en tant qu'expert judiciaire) ont cependant été sollicités par la Justice pendant encore de trop longues années[15], faute d'un recoupement organisé des informations entre magistrats, tribunaux ou cours d'appel.

6 De la prise de position à l'action scientifique.

La motion mettait l'accent sur la volonté de collaboration des chercheurs académiques. Elle traçait les pistes de cette collaboration : accroître les connaissances, construire des corpus, proposer des procédures rigoureuses d'évaluation et de certification des méthodes, techniques et experts. Les chercheurs ont travaillé sur les premiers thèmes, malgré la faiblesse du soutien institutionnel pour des travaux concernant des applications judiciaires, en s'appuyant sur le développement de la reconnaissance du locuteur et de la biométrie (se référer par exemple[16] à Besacier 2000 ; Gravier 1997 ; Magrin-Chagnolleau 2000 ; Perrot 2007, 2008). Au fil des années, des travaux dédiés au contexte judiciaire ont été proposés, d'abord hors de France (Champod 2000 ; Meuwly 2001, 2003) puis en France (Ajili 2016a, 2016b, 2017 ; Kahn 2010a, 2010b, 2011a, 2011b). Un premier projet ANR dédié à la fiabilité des méthodes[17] a pris place en 2013. Le projet VoxCrim[18] lui a succédé en

[13] Par opposition à l'intime conviction de l'expert, fusse-t-elle basée sur une solide expérience et/ou un cursus scientifique.

[14] La Justice est déjà une machine lente. De plus les avocats de la défense ont tendance à contacter l'AFCP au moment du procès en appel, qui intervient bien après l'expertise initiale.

[15] Voir par exemple l'affaire Willy Bardon / Élodie Kulik en 2019.

[16] Seuls quelques exemples parmi les plus proches du contexte judiciaire ont été choisis.

[17] ANR-12-BS03-0011 FABIOLE (LIA[20], LNE[21] et LIG, Univ. Grenoble Alples & CNRS).

2017. Ce projet associe pour la première fois les deux laboratoires nationaux de police scientifique français, le SCPTS et l'IRCGN[19], au LIA, LPL et LPP[20], trois laboratoires académiques, et au laboratoire national de métrologie français, le LNE[21]. VoxCrim est dédié à la comparaison de voix appliquée au domaine criminalistique et propose tout à la fois de travailler sur le chapitre des connaissances fondamentales, sur une procédure pratique d'accréditation des méthodes de comparaison de voix dans le domaine criminalistique et sur la formation des acteurs impliqués.

7 La relation aux médias

Durant toutes ces années, la relation aux médias a été une des questions difficiles. Dans la plupart des cas, les médias s'intéressent à la question « à chaud », soit quand une affaire défraie la chronique. Souvent, les délais sont très courts et les journalistes souhaitent des éléments tranchés et démonstratifs, particulièrement quand la presse télévisuelle est concernée. Expliquer une action sur la durée, appuyée sur une motion proposée il y a 30 ans, n'est déjà pas simple en soi. De plus, les journalistes contactent les chercheurs rencontrés précédemment ou utilisent des contacts institutionnels, sans forcément porter attention au thème de leur requête (*traitement du langage* et/ou *IA* ou même *informatique* représentent la granularité souvent retenue). Les chercheurs académiques (et leurs institutions) ont tendance à répondre aux demandes même lorsqu'elles sont à la frontière de leur zone de confort. Malheureusement, ce type d'échanges avec les médias a souvent produit des éléments manquant de précisions ou même faux, ce qui n'est pas sans conséquences quand ils reviennent dans les prétoires, ce qui arrive fréquemment. Nous rencontrons aussi des *professionnels des médias* cherchant non pas à informer mais seulement à appuyer une opinion préétablie, à faire du sensationnel ou à remplir coûte que coûte un espace. Dans ce cas, il est arrivé que le message diffusé soit diamétralement opposé à ce que le chercheur souhaitait exprimer.

8 Limites du témoignage scientifique

Comme vu précédemment, la légitimité des représentants de l'AFCP dans les tribunaux repose largement sur le cadre strict de la démarche scientifique objective qu'ils ont choisi. Mais les questions qui peuvent leur être posées dépassent souvent ce cadre. Par exemple, au cours d'un témoignage récent[22], juges, procureur et avocats ont demandé au témoin scientifique si une personne est capable d'identifier une voix à l'oreille, dans quelles conditions et avec quelle fiabilité, si tous sont égaux devant cette tâche et, enfin, si nous reconnaissons mieux un familier qu'un individu lambda. De manière évidente, les réponses pouvaient peser lourdement sur le verdict et les attentes du tribunal étaient fortes. Lorsque, comme ici, il n'y a pas suffisamment de travaux scientifiques probants publiés pour répondre en bénéficiant de la « garantie » scientifique, le témoin scientifique doit se restreindre à dire uniquement cela ; ce qui s'est avéré difficile dans l'exemple cité précédemment, car le besoin d'information exprimé par les intervenants était prégnant, et parce que

---

18 VoxCrim, ANR-17-CE39-0016.

19 Le service Central de la Police Technique et Scientifique (SCPTS) de la Police Nationale et l'Institut de Recherche Criminelle de la Gendarmerie Nationale (IRCGN).

20 Le Laboratoire Informatique d'Avignon (LIA), Avignon Univ., le Laboratoire Parole et Langage (LPL), Aix-Marseille Univ. & CNRS et le Laboratoire de Phonétique et Phonologie (LPP), Univ. Sorbonne Nouvelle & CNRS.

21 Le Laboratoire National de métrologie et d'Essais (LNE).

22 L'affaire Kulik/Bardon, déjà citée.

l'expérience personnelle peut pousser un chercheur à penser que son opinion, voire son intuition, peuvent être pertinentes et porteuses d'informations utiles, à juste titre ou non.

9 Conclusion… ou bilan d'étape ?

Il aura fallu l'affaire d'Outreau pour que le rôle de la science dans les tribunaux soit reconnu comme une question sociétale majeure en France, tant le risque lié à un mauvais usage est apparu important à cette occasion. Dans le cadre de l'identification de voix en milieu judiciaire, bien avant cette affaire, les scientifiques francophones de la communication parlée ont dès 1990 pris conscience de cela et se sont attachés à présenter une position claire, définie par une motion unanimement soutenue et largement diffusée. La motion de 1990 a servi de référentiel durant les trois décennies passées, sans jamais être modifiée ou amendée, ce qui peut paraître étonnant au regard des avancées réalisées durant la même période dans le domaine du traitement automatique de la parole notamment. Loin de marquer un immobilisme ou un conservatisme, ce résultat vient de la structure de la motion, appuyée sur trois piliers : une prise de conscience de la spécificité de l'expertise scientifique judiciaire en identification de voix, la nécessité d'un moratoire tant que le domaine et les experts n'ont pas fait la preuve de leur maturité scientifique et une ouverture vers l'avenir, à la fois pour développer les connaissances scientifiques sur l'identification de voix et pour mettre en place des coopérations avec les acteurs de la police scientifique comme de la Justice.

En ce qui concerne le premier pilier, l'action sur la durée des chercheurs de l'AFCP a sans conteste créé une prise de conscience en France, dans les laboratoires académiques, la police scientifique et les tribunaux. Cette prise de conscience a dépassé le cadre francophone et a été étendue au niveau international (Bonastre 2003, 2004 ; Campbell 2009). La question des applications criminalistiques de la reconnaissance du locuteur, quasi ignorée jusqu'alors par le milieu académique, est devenue incontournable (voir le groupe de l'ISCA *SPLC*[23], Bonastre 1997, les workshops *Speaker Odyssey* ou les *Special Event on Speaker Comparison for Forensic and Investigative Applications* des conférences Interspeech 2015, 2016 et 2017).

La demande (la motion) puis l'injonction (la pétition) d'un arrêt des expertises judiciaires en identification de voix, le deuxième pilier de la motion, a été plus long à se mettre en place. Les premières actions ont bien permis le respect de ce moratoire de la part des scientifiques de l'AFCP (à de très rares exceptions près), mais les laboratoires de police scientifique et des acteurs privés ont continué les expertises. Cela a amené les représentants de l'AFCP à se présenter dans les tribunaux comme témoins scientifiques, ou *sachants*, pour expliquer la position de l'AFCP et les limites de l'expertise judiciaire en identification de voix. Ce mode d'action est devenu central quand ces représentants ont été confrontés à un phénomène de charlatanisme. Estimer l'influence de ces interventions dans les tribunaux est salutaire mais difficile à réaliser. Les verdicts ne sont pas appropriés pour cela, car nous ne cherchons pas à influencer la Justice mais à l'informer, sans compter que l'importance des expertises en identification de voix est variable suivant les procès. Le meilleur outil disponible est la lecture des attendus et de la presse. Cette lecture montre que cette action a permis à la magistrature de mieux juger de la force de conviction à attendre des expertises en identification de voix, ainsi que des nombreuses limites de celles-ci, tant au niveau scientifique qu'en termes de mise en œuvre. Il apparaît également clairement que les actions entreprises ont permis de mettre en évidence puis de réduire le phénomène de charlatanisme, et c'est là un résultat majeur pour l'ensemble des justiciables concernés. Si elle était peu valorisante au départ, la perception des témoins scientifiques de l'AFCP par les tribunaux s'est nettement améliorée au fil des années. Le point clé a été la promotion de la méthodologie scientifique, avec ses garanties

[23] Special Interest Group on SPeaker and Language Characterization (SIG/SPLC)

d'objectivité, tout en se gardant de dévaloriser l'intime conviction d'un expert, certes subjective et n'offrant donc pas les mêmes garanties, mais reposant sur une expérience personnelle pouvant être porteuse de sens. Pour un témoin scientifique, rester sur cette position d'objectivité scientifique quand les questions débordent de ce champ n'est pas simple mais est primordial : se restreindre à dire que la science, donc vous, ne sait pas répondre quand elle n'a pas assez avancé sur un sujet est éminemment préférable plutôt qu'énoncer une position basée sur votre intime conviction.

Quant au troisième pilier, la coopération scientifique, si les chercheurs académiques ont pu très tôt travailler de manière détournée sur l'identification judiciaire de la voix, jusqu'à ces dernières années cela a été peu soutenu et sans participation active des laboratoires de police scientifique. L'engagement permanent de l'AFCP a permis de créer progressivement, étape par étape, le contexte propice à un projet regroupant laboratoires de police scientifique et laboratoires académiques sur des objectifs clairs et respectueux de la motion de 1990, comme illustré par le projet ANR VoxCrim. Outre le fait qu'il associe laboratoires de recherche publique, de police scientifique et de métrologie, VoxCrim s'éloigne de la course à la performance pour se recentrer sur la fiabilité des méthodes et les solutions pour attester de celle-ci, dans la lignée de la motion de 1990.

Cet exemple illustre une évolution dans la lecture de la motion. Celle-ci était perçue initialement comme une injonction formelle à ne pas participer au monde de l'expertise judiciaire en identification de voix tant qu'une approche ne faisant pas d'erreur n'avait pas été proposée et validée, hypothèse qui apparaissait peu probable. En 2020, le même texte prend un sens plus proche de son message principal : une méthode ne peut être utilisée pour une expertise judiciaire donnée que si, pour ce cas même, la fiabilité est évaluée et attestée au préalable. Ce qui revient à dire que la marge d'erreur doit être connue et attestée au cas par cas avant de réaliser une expertise, à charge pour le tribunal d'évaluer l'intérêt de celle-ci en fonction de l'affaire et de la marge d'erreur estimée. Ce n'est pas en effet au chercheur académique de déterminer si une information, aussi minime soit-elle, est pertinente ou pas dans le contexte judiciaire : son rôle se limite à définir si cette information repose sur une base scientifique solide ou non et, le cas échéant, à aider le tribunal à mieux évaluer la pertinence de l'information proposée.

Dans ce texte, il n'a encore jamais été fait mention de différences entre une méthode dite *experte* (souvent appelée *phonétique*) et une méthode dite *automatique*. Cela est volontaire car pour se prévaloir de la qualité de *scientifique*, celles-ci partagent les mêmes nécessités d'évaluation et d'objectivité (Bonastre 2008 ; Campbell 2009 ; Rose 2006). Les modalités de mise en œuvre de ces deux approches, comme les contraintes et limites associées, seront évidemment différentes, mais cette question sort du champ de cet article.

La relation aux médias reste un point mal traité, majoritairement laissé à l'initiative des médias à l'occasion d'affaires judiciaires fortement médiatisées. Pour conserver au message toute sa teneur, il est essentiel que chaque mot reproduit dans les médias soit pesé précautionneusement et reste en conformité avec la position de l'AFCP, en gardant à l'esprit que chaque communication pourra être citée pendant de longues années dans les tribunaux. Recentrer la relation média sur un ou quelques porte-paroles apparaît comme une solution souhaitable pour éviter une dissipation du message.

Enfin, vouloir agir dans ce domaine implique une dose certaine de patience : cet article relate trente années d'actions et ressemble pourtant à un bilan d'étape provisoire. La Justice ne sait pas s'arrêter aux positions générales, c'est le lieu de l'étude des cas particuliers et son tempo est lent. Une position pourtant bien reconnue peut ainsi mettre plusieurs années pour s'imposer. Il faut également accepter la frustration d'être confronté aux mêmes limitations du système législatif/judiciaire pendant parfois des dizaines d'années comme, par exemple, le fait que les experts réalisant des expertises en identification de voix en France soient inscrits dans la catégorie « acoustique ».

Remerciements

Merci tout d'abord à l'AFCP et ses précurseurs, à tous leurs comités et présidents successifs depuis 1990 qui ont soutenu sans faille cette action. Sans l'engagement personnel permanent de Louis-Jean Boë, celle-ci n'aurait ni grandi ni perduré sur la durée, qu'il en soit remercié ici. Merci à Gilles Adda et Karen Fort, pour avoir sollicité cet article (et avoir su à la fois insister et soutenir l'auteur dans ses moments de doute). Merci à l'ensemble des relecteurs et conseillers, avec une mention spéciale pour Nathalie Vallée et Emmanuel Ferragne.

Références

AJILI, M., BONASTRE, J. F., KAHN, J., ROSSATO, S., & BERNARD, G. (2016a, May). Fabiole, a speech database for forensic speaker comparison. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (p. 726-733).

AJILI, M., BONASTRE, J. F., KHEDER, W. B., ROSSATO, S., & KAHN, J. (2016b, December). Phonetic content impact on forensic voice comparison. In 2016 IEEE Spoken Language Technology Workshop (SLT) (p. 210-217). IEEE.

AJILI, M., BONASTRE, J. F., KHEDER, W. B., ROSSATO, S., & KAHN, J. (2017). Homogeneity Measure Impact on Target and Non-Target Trials in Forensic Voice Comparison. In INTERSPEECH (p. 2844-2848).

BESACIER, L., BONASTRE, J. F., & FREDOUILLE, C. (2000). Localization and selection of speaker-specific information with statistical modeling. Speech Communication, **31**(2-3), (p. 89-106).

BOË, L. J., BIMBOT, F., BONASTRE, J. F., & DUPONT, P. (1999). De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. Langues, **2**(4), (p. 270-288).

BOË, L. J. (2000). Forensic voice identification in France. Speech Communication, **31**(2-3), (p. 205-224).

BOË, L. J., & BONASTRE, J. F. (2012a, June). L'identification du locuteur : 20 ans de témoignage dans les cours de Justice. Le cas du LIPSADON « laboratoire indépendant de police scientifique ». In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP (p. 417-424).

BOË, L. J. et BONASTRE, J.F. (2012b). Expertise de la voix : identifier le locuteur à partir d'écoutes téléphoniques ? Des expertises à la recherche d'une caution scientifique… ou le cas du laboratoire Lipsadon, J'essaime, numéros 22 et 23, 2012.

BONASTRE, J. F., BIMBOT, F., BOË, L. J., Campbell, J. P., Reynolds, D. A., & Magrin-Chagnolleau, I. (2003). Person authentication by voice: A need for caution. In Eighth European Conference on Speech Communication and Technology.

BONASTRE, J. F., BIMBOT, F., BOË, L. J., Campbell, J. P., Reynolds, D. A., & Magrin-Chagnolleau, I. (2004). Authentification des personnes par leur voix : un nécessaire devoir de précaution. Journées d'Etudes de la Parole, (p. 33-36).

BONASTRE, J. F., & MATROUF, D. (2008). La reconnaissance du locuteur : un problème résolu ? Journées d'études sur la Parole (JEP).

BONASTRE, J. F., MAGRIN-CHAGNOLLEAU, I., EULER, S., PELLEGRINO, F., ANDRÉ-OBRECHT, R., MASON, J. S., & BIMBOT, F. (2001). SPeaker and Language Characterization (SpLC): A Special Interest Group (SIG) of ISCA. In Seventh European Conference on Speech Communication and Technology.

CAMPBELL, J. P., SHEN, W., CAMPBELL, W. M., SCHWARTZ, R., BONASTRE, J. F., & MATROUF, D. (2009). Forensic speaker recognition. IEEE Signal Processing Magazine, **26**(2), (p. 95-103).

CHAMPOD, C., & MEUWLY, D. (2000). The inference of identity in forensic speaker recognition. Speech communication, **31**(2-3), (p. 193-203).

GRAVIER, G., MOKBEL, C., & CHOLLET, G. (1997). Model dependent spectral representations for speaker recognition. In Fifth European Conference on Speech Communication and Technology.
KAHN, J., ROSSATO, S., & BONASTRE, J. F. (2010a, March). Beyond doddington menagerie, a first step towards. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (p. 4534-4537). IEEE.
KAHN, J., AUDIBERT, N., ROSSATO, S., & BONASTRE, J. F. (2010b). Intra-speaker variability effects on Speaker Verification performance. In Odyssey (p. 21).
KAHN, J., AUDIBERT, N., ROSSATO, S., & BONASTRE, J. F. (2011a). Speaker verification by inexperienced and experienced listeners vs. speaker verification system. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (p. 5912-5915). IEEE.
KAHN, J. (2011b). Parole de locuteur: performance et confiance en identification biométrique vocale. Thèse de doctorat.
MAGRIN-CHAGNOLLEAU, I., GRAVIER, G., SECK, M., BOEFFARD, O., BLOUET, R., & BIMBOT, F. (2000). A further investigation on speech features for speaker characterization. In Sixth International Conference on Spoken Language Processing.
MEUWLY, D., & DRYGAJLO, A. (2001). Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). In 2001: A Speaker Odyssey-The Speaker Recognition Workshop.
MEUWLY, D., GOODE, A., DRYGAJLO, A., GONZALEZ-RODRIGUEZ, J., & MOLINA, J. L. (2003, September). Validation of forensic automatic speaker recognition systems: Evaluation frameworks for intelligence and evidential purposes. In Forensic Science International (Vol. 136, pp. 364-364).
Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. Computer Speech & Language, **20**(2-3), (p. 159-191).
PERROT, P., AVERSANO, G., & CHOLLET, G. (2007). Voice disguise and automatic detection: review and perspectives. In Progress in nonlinear speech processing (p. 101-117). Springer, Berlin, Heidelberg.
PERROT, P., & CHOLLET, G. (2008). The question of disguised voice. Journal of the Acoustical Society of America, **123**(5), 3878.